ISSN 2466-135X
Vol.7, No.1

The 7th International Conference on

BIG DATA APPLICATIONS AND SERVICES (BIGDAS2019)

Proceeding

August 21–24, 2019
Jeju Island, South Korea

Hosted by
Korea Big Data Service Society

THE KOREA
BIG DATA SERVICE SOCIETY
한국빅데이터서비스학회

제주대학교
JEJU NATIONAL UNIVERSITY
SW융합교육센터
Software Convergence Education Center

충북대학교
빅데이터연구소

Table of Contents

Improvement of Cross-Modal Retrieval Performance for Visual-Semantic Data Through Adversarial Learning

Sanghyuck Na[1], Daeung Kim[1], Sanghyun Seo[1] and Juntae Kim[1]

[1] Department of Computer Engineering, Dongguk University
{shna, dukim, shseo, jkim}@dongguk.edu

Abstract. Cross-modal retrieval is a task to retrieve the most similar target data which has different modality for a query, and it can be achieved through mapping multi-modal data into common vector space. Recently, various methods of cross-modal retrieval between image and text have been proposed by using visual-semantic embedding with deep neural network. However, since most of existing methods use only a set of distance metric as loss functions, there are some problems that the characteristic of the original modality remains in the embedded vector, which degrades the performance of cross-modal retrieval. In this paper, we propose a method to improve the cross-modal retrieval performance by applying the adversarial learning method. The proposed adversarial learning method makes the discriminator model to try to distinguish the original modalities of the embedded vectors and the embedding model to try to reduce the differences of modality characteristic between embedded vectors. The experimental results show that the proposed method has better cross-modal retrieval performance than the embedding model only using distance metric as loss functions, based on general deep neural network.

Keywords: Cross-modal Retrieval, Visual-Semantic Embedding, Adversarial Learning, Transfer Learning

1 Introduction

Cross-modal retrieval is a task to retrieve the most similar target data which has different modality for query. One of the various methods for conducting cross-modal retrieval is the method using the embedding model [1]. Comparison of multi-modal data is considered a difficult task due to the dimension difference. However, the task is handled efficiently by using common embedding, since the high-dimensional data is embedded into lower-dimensional common vector space in the embedding model. We can also reduce the computation in retrieving target data by reducing the dimension of multi-modal data through embedding model. Recently, researches of visual-semantic embedding using image and text have attracted attention [2-5]. However, traditional methods of visual-semantic embedding use independent embedding model for each modality, so there is a weakness that the characteristic of the modality in embedded vectors has a bad influence on cross-modal retrieval [6].

In this paper, we propose the adversarial learning method in which multi-modal data is embedded by semantic. In this the adversarial learning method, the embedding model tries to deceive the discriminator model, and the discriminator model tries to distinguish the original modality of the embedded vectors. To deceive the discriminator model, the embedding model tries to reduce the modality characteristic differences between embedded vectors while the discriminator model tries to distinguish the original modality of the embedded vector. Therefore, distinguishing modality characteristic become difficult between embedded data in common vector space, and it can improve cross-modal retrieval performance.

2 Related Works

Unlike simple classifiers, cross-modal retrieval refers to a method for retrieving targets such as text, image, for the input query. If we can embed the data of different type which has different dimension and characteristic into common vector space, it is possible to semantically compare multi-modal data. For example, cross-modal retrieval method using cosine similarity between visual-semantic data has been proposed [4-5].

The visual-semantic embedding is a general method of cross-modal retrieval. In this work, there are embedding both visual data such as image feature vector and semantic data such as distributed vector obtained from language model into common vector space. The embedding model learns the relationship between visual features of image vector and the semantic features of word vector. So, it is possible to recognize the data not included in the training set unlike a simple classifier [2].

There has been a method in which an embedded image vector is set as a query and the embedding model is trained so that text vector and audio vector having the same meaning with query are embedded closer to the embedded image vector. By minimizing the distances between text, audio, and image which are semantically similar, this method shows that text and audio were closely aligned around the image in common vector space [6]. However, this method does not effectively reduce the distance between the modalities, except for the image, as it only makes the distance from the image to the other two modalities relatively close to each other than the distance between each vector, audio and text.

Another method is to use the adversarial learning method for multi-modal data [7]. In this work, there are image and text intra-modality discriminator and inter-modality discriminator. By using the inter-modality discriminator which distinguishes the image common representation feature from the text common representation feature, cross-modal retrieval performance is improved. In a similar approach, there is a method using the Siamese network and adversarial learning method for multi-modal data [8]. The discriminator distinguishes the image special feature from the image networks and the text special feature from the text networks. In addition, the Siamese network are trained in embedding visual-semantic data to consider the semantically similarity between each similar image feature vector and text feature vector, which makes the performance of cross-modal retrieval to be improved.

3 Visual-Semantic Embedding with Adversarial Learning

3.1 Visual-Semantic Embedding with Modality Discriminator

To improve cross-modal retrieval performance, visual-semantic embedding technique is used. In visual-semantic embedding, a pair semantic data with similar meaning is embedded relatively close to each other, where distinct pair is embedded relatively distant. In addition, image and text data within a class are considered semantically similar and placed closely. In general, there is a tendency to perform visual-semantic embedding using mean squared error (MSE) and triplet margin loss (TRP) [4-5]. These loss functions have some problems that the characteristic of the original modality remains in the embedded vector from embedding model, which degrades the performance of cross-modal retrieval [6].

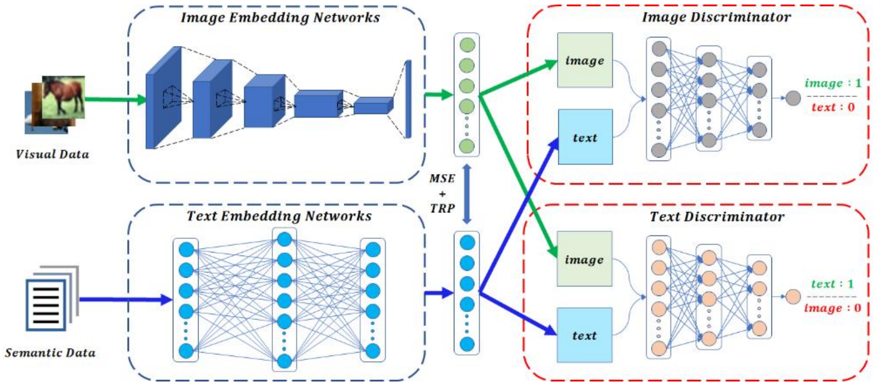


Fig. 1. An overview of visual-semantic embedding model with proposed adversarial learning method

In this paper, we propose the adversarial learning (AL) method in embedding visual-semantic data for cross-modal retrieval. The above Fig. 1. is the model architecture of the proposed method. The proposed adversarial learning method makes the image discriminator model that regards image as a real and text as a fake to distinguish the modalities of image and text [9]. At that time, the text embedding networks learns to deceive the image discriminator model. This adversarial learning is done in the same way for a text discriminator model which regards text as a real data and image as a fake data. Through these a method, learning is performed so that image and text are embedded in a similar location to each pair data, and the embedding model is trained so that the specific modality characteristics of the original data are not distinguished, there by contributing to improvement of the cross-modal retrieval performance.

The image feature vector extracted from image embedding networks and text feature vector extracted from the text embedding networks are embedded in common vector space. The image discriminator model and the text discriminator model make the image and text feature vectors embedded in the common vector space not to be distinguished. As a result, each data lost its features about modality, so regardless of the type of modality semantically similar data is embedded close to each other.

3.2 Loss Function for Embedding Model and Modality Discriminator

Our goal is to embed visual and semantic data in the common vector space without influence of modality. To achieve this goal, we use two methods. The first method is the general visual-semantic embedding method. By using mean squared error and triplet loss function, a pair of visual data and semantic data having the same meaning get closer to each other while semantic data having the different meaning get farther away. The second is adversarial learning method. This method regards embedding model as generator and attach the discriminator so that the gap is reduced between the distribution of visual and semantic data.

3.2.1 Embedding Loss

Let v be a visual data, s be a semantic data and $f(v)$ be the embedded vector of visual data v by the image embedding networks f. And $g(s)$ be the embedded vector of semantic data s by the text embedding networks g.

To optimize embedding model, we use mean squared error and triplet loss function using L_2 norm as distance metric. The equation (1) is L_2 norm means the distance between the embedded vector $f(v^+)$ and $g(s^+)$ having the same meaning. The equation (2) is the mean squared error between $f(v^+)$ and $g(s^+)$. By minimizing this loss function, $f(v^+)$ and $g(s^+)$ are getting closer to each other. The equation (3) is the triplet loss function. We consider $f(v^+)$ to be an anchor and a positive image sample and $g(s^+)$ to be positive text sample, and $g(s^-)$ to be negative text sample. By minimizing this loss function, the distance from the $f(v^+)$ to $g(s^+)$ is to be closer than distance from the $f(v^+)$ to $g(s^-)$ by at least margin α.

$$d(v^+, s^+) = \|f(v^+) - g(s^+)\|_2 \ . \tag{1}$$

$$MSE = \frac{1}{n}\sum d(v^+, s^+)\,, where\ n = batch\ size\ . \tag{2}$$

$$TRP = \sum_{n} max(\alpha + d(v^+, s^+) - d(v^+, s^-), 0)\ . \tag{3}$$

In equations (2) and (3) n denotes batch size. Since we use discriminator too, the other loss term (so called generator loss in GAN [9]) is added to the embedding loss function. We will describe it together with adversarial learning loss at next section.

3.2.2 Adversarial Learning Loss

Let the image discriminator model be D_v, value function be V To optimize image discriminator model, we use the value function $V(D_v)$ as the loss function. In the equations (4) and (5), the visual data $f(v)$ is considered to be a real and the semantic data $g(s)$ is considered to be a fake, and the text discriminator model D_v is learned to discriminate what modality the input vector originated from (image or text) and text embedding networks learns how to embed semantic vector in a form similar to visual vector so that image discriminator model can't distinguish what modality the input vector come from. The equation (5), which is included in the equation (4), has a relationship with embedding model. By minimizing this term, the embedding model learn how to embed a semantic data in the form similar to visual data so that the image discriminator model cannot distinguish the modality of the data. It is vice versa to the value function of text discriminator model $V(D_s)$.

$$\underset{D_v}{max}\, V(D_v) = \mathbb{E}_{v^+ \sim p_v(v)}\left[\log D_v\big(f(v^+)\big)\right] + \mathbb{E}_{s^+ \sim p_s(s)}[\log(1 - (D_v(g(s^+))))]\,. \quad (4)$$

$$\min_{g} V(g(s^+)) \;=\; \mathbb{E}_{s^+ \sim p_s(s)}[\log(1 - (D_v(g(s^+))))]\,. \quad (5)$$

3.2.3 Total Embedding Loss with Adversarial Learning

The equations (6) and (7) summarizes the loss functions of our method. By optimizing the terms of MSE, the both image and text embedding networks learn to make both the $f(v^+)$ and $g(s^+)$ to get closer to each other, and optimizing the terms of TRP, both model learn to make the distance between $f(v^+)$ with $g(s^+)$ to be closer than the distance between $f(v^+)$ with $g(s^-)$.

The image discriminator model D_v regards the embedded vector from the image modality $f(v^+)$ as a real and learns the embedded vector from text modality $g(s^+)$ as a fake. By the terms of $V(f(v^+))$, D_v learns to distinguish $f(v^+)$ and $g(s^+)$ and the text embedding networks learns how to embed a semantic data in the form similar to visual data so that image discriminator model cannot distinguish where the input data come from. It is vice versa to the text discriminator model D_s and image embedding networks do.

$$L_f = MSE + TRP + V(f(v^+))\,. \quad (6)$$

$$L_g = MSE + TRP + V(g(s^+))\,. \quad (7)$$

Where L_f, L_g denote the loss function of $f,\ g$. By minimizing them, the gap is reduced between the distribution of visual data and semantic data.

4.1 Experiment Setup

We performed the experiments on datasets, CIFAR-100, STL-10 [12, 13]. The CIFAR-100 is 32x32 RGB color image and consists of 20 super classes and 5 subclasses for each superclass. STL-10 is 96x96 RGB color image, consists of 10 classes, 5,000 train set and 1,000 test set. The semantic data is obtained from pretrained word vector with 300-dimensional obtained by google news dataset using word2vec [10].

For each embedding model, we conducted the image to text cross-modal retrieval task and measure *hit@k*. The *hit@k* is one of the usual evaluation metrics in cross-modal retrieval fields. It means the percentage of the queries containing true label among the top *k* retrieval results.

There are five cases of the experimental setup. The first case is the image classifier that modified VGG19 [11]. The second case is the baseline model, which uses the MSE and TRP loss functions only. The third case is the baseline model applying transfer learning [14]. We use pre-trained weights except the last FC layer in the classification model. The fourth case is the proposed model using adversarial learning method based on baseline embedding model. The fifth case is the proposed model using both the pre-trained weight of classifier and adversarial learning method based on baseline embedding model.

4.2 Training Detail

The image embedding networks follows the standard VGG 19, except the input layer and the output layer [11]. The input is 32x32 RGB image and the output is the 300-dimensional vector. The text embedding networks consists of three FC layers. The size of each layer is 512, 512, 300 separately.

The adversarial learning network consist of two networks, the one is about Image modality and the other is about text modality. Each of them has three FC layers. And the size of each layer is 128, 64, 1. As the activation function, LeakyReLU with a negative slope of 0.02 is applied to the first and second FC layer, and sigmoid function is applied to third FC layer. Batch normalization is used behind the second FC layer as a Regularization technique. In the experiment 2~5, we optimized image and text embedding networks with the MSE and TRP loss function, using SGD optimizer with learning rate 0.01.

In the experiment on STL-10 dataset, we resized the input image of 96x96 pixel to 32x32. And optimized adversarial learning network using the Adam optimizer. The hyper-parameters of Adam are β_1 = 0.5, β_2 = 0.999, and learning rate 0.0003. In the experiment 2~5 on CIFAR-100 dataset, we optimized adversarial learning network using the Adam optimizer. The hyper-parameters of Adam are β_1 = 0.5, β_2 = 0.999, and learning rate 0.0002. The scheduler of the Adam optimizer reduced the learning rate, if the value of the previous loss function does not decrease over 2 epochs. In all experiments on STL-10, CIFAR-100 dataset, which has trained for 30 epochs.

4.3 Experiment Result and Discussion

The results of experiments are described in Table 1. Firstly, we compared our models with the image classification model. On the STL-10, the baseline showed lower score at *hit@1* task than the classifier. But the Baseline + Transfer + AL (ours) model outperformed on the *hit@1*, *hit@5* task than the baseline model. On the CIFAR-100, the Baseline + Transfer + AL (ours) model achieved highest *hit@1* and *hit@5* score.

Table 1. *hit@k* on STL-10 and CIFAR-100

Dataset	Model	*hit @ 1*	*hit @ 5*
STL-10	Classification	0.6735	-
	Baseline (embedding model only)	0.4577	0.9250
	Baseline + Transfer	0.6909	0.9377
	Baseline + AL (ours)	0.4985	0.9185
	Baseline + Transfer + AL (ours)	**0.6953**	**0.9500**
CIFAR -100	Classification	0.5898	-
	Baseline (embedding model only)	0.2617	0.6199
	Baseline + Transfer	0.6770	0.8609
	Baseline + AL (ours)	0.2608	0.6300
	Baseline + Transfer + AL (ours)	**0.6800**	**0.8676**

Secondly, we examined the performance of *hit@1, hit@5* can be affected by transfer learning. The models with transfer learning outperformed than non-transfer learning models. As can be seen from Table 1, the models with transfer learning achieved the performance improvement on *hit@1* task at least 20% up to 41% (Baseline + Transfer + AL on STL-10, Baseline + Transfer + AL on CIFAR-100) and *hit@5* at least 1% up to 23% (Baseline + Transfer on STL-10, Baseline + Transfer + AL on CIFAR-100).

Thirdly, we examined the performance of *hit@1, hit@5* can be affected by adversarial learning. The models with adversarial learning outperformed than the model non-adversarial learning models. The models with adversarial learning achieved the performance improvement on *hit@1, hit@5* task.

STL-10, CIFAR-100 has relatively a few data for one class compared to other datasets (e.g., CIFAR-10). However, our method can effectively improve the performance of cross-modal retrieval.

5 Conclusions

We have proposed the method of the cross-modal retrieval performance improvement for visual-semantic data through adversarial learning. The proposed method trains the embedding model so that the characteristic of the original data modality not to be distinguished by using discriminator model. The proposed method shows that the cross-modal retrieval performance can be improved by removing the inter-modality characteristic of the original data.

In the future work, we plan to design more sophisticated loss functions, acquire various multi-modal datasets, and perform cross-modal retrieval for large-scale data.

6 Acknowledgments

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7083279)

References

1. Kang, C., SC, L., Li, Z., Cao, Z., & Xiong, G.:Learning Deep Semantic Embeddings for Cross-Modal Retrieval. In: JMLR: Workshop and Conference Proceedings 80:1–16. ACML, New Zealand (2017)
2. Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., & Mikolov, T.:Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems, pp. 2121-2129. NIPS, Lake Tahoe (2013)
3. Guo, G., Zhai, S., Yuan, F., Liu, Y., & Wang, X.: Vse-ens.: Visual-semantic embeddings with efficient negative sampling. In: Thirty-Second AAAI Conference on Artificial Intelligence. AAAI Hilton New Orleans Riverside (2018)
4. Cao, Y., Long, M., Wang, J., Yang, Q., & Yu, P. S.: Deep visual-semantic hashing for cross-modal retrieval. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1445-1454. ACM.
5. Hahn, M., Silva, A., & Rehg, J. M. In.: Action2Vec.: A Crossmodal Embedding Approach to Action Learning. arXiv preprint arXiv:1901.00484 (2019)
6. Aytar, Y., Vondrick, C., & Torralba, A. In.: See, hear, and read: Deep aligned representations. arXiv preprint arXiv:1706.00932 (2017)
7. Peng, Y., & Qi, J.: Cm-gans: Cross-modal generative adversarial networks for common representation learning. In: ACM Transactions on Multimedia Computing, Communications, and Applications. TOMM, New York
8. Wen, X., Han, Z., Yin, X., & Liu, Y. S.: Adversarial Cross-Modal Retrieval via Learning and Transferring Single-Modal Similarities. arXiv preprint arXiv:1904.08042 (2019)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y.: Generative adversarial nets. In: In Advances in neural information processing systems, pp. 2672-2680. NIPS, CANADA (2014)
10. Mikolov, T., Le, Q. V., & Sutskever, I.: Exploiting similarities among languages for machine translation arXiv preprint arXiv:1309.4168 (2011)
11. Simonyan, K., & Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
12. Krizhevsky, A., & Hinton, G.: Learning multiple layers of features from tiny images (Vol. 1, No. 4, p. 7). Technical report, University of Toronto (2019)
13. Coates, A., Ng, A., & Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp. 215-223. USA (2011)
14. Pan, S. J., &Yang, Q.:Devise: A survey on transfer learning. In:IEEE transations on knowledge and data engineering, pp. 1345-1359. TKDE (2009)

Pipeline Refinement and Performance Analysis for Shading Based Sunlight Amount Analyze System: Fully Parallel Approach with Aid of Modern GPU

Woosuk Shin[1], Nakhoon baek[2,*]

[1,2] Department of Computer Science and Engineering,
Kyungpook National Universirty,
41566 Daegu, Korea
mell03@naver.com[1] , oceancru@gmail.com[2,*]

Abstract. In modern architecture, analyzing sunlight interference to existing environment is important factor for constructing new architecture. Thus, CAD systems provide utility to analyze sunlight reachability for selected surface. However, the system utilizes ray-tracing-like algorithm to trace reachability for given surface and algorithm itself consumes long computation time. The biggest disadvantage of using such algorithm in system is that it constraints number of polling point per surveying area (i.e., one polling point per four square meter) to minimize computation. To improve performance and remove polling point problem of existing sunlight analysis system, we already proposed shading based system. In this paper, we introduce more optimized pipeline design for the idea to give fully parallelism to the system which utilizes modern GPU. With our optimized design, we could achieve up to three times faster simulation results depending on number of simultaneous running multiple-contexts compared to well-known utility, achieving almost identical result.

Keywords: Sunlight Duration Analysis, Massively Parallel Processing, Big Data Generation, Big Data Analysis, Big Data Processing,GPGPU

1 Introduction

In modern architecture, analyzing sunlight interference to existing environment is important factor for constructing new architecture [1]. For example, most countries regulates that certain time of sunlight reachability should be guaranteed for residential area. Thus, to aid calculation of total amount of sunlight for certain area, CAS systems provide utility to analyze sunlight reachability for selected surface. However, these systems which utilizes ray-tracing-like algorithm [2] to trace reachability for given surface for each single step and ray-tracing algorithm itself consumes long computation time. Thus, systems using ray-tracing-like algorithm demands user to select several surfaces and density of polling point per surveying area to minimize computation process. However, even with small polling points, as selected surface increases, amount of computation increases rapidly. To overcome overall limitations, we already designed a system to analyze sunlight duration using shading.

1.1 Sunlight Duration Analyze System using Shading

In [3], we introduced system design to analyze sunlight duration using accumulative shadow map, which is based on a shading effect of light due to nearby architectures, and adopted it to urban scenario [4] which consists of high-story buildings. Fig. 1 is overall pipeline of the system.

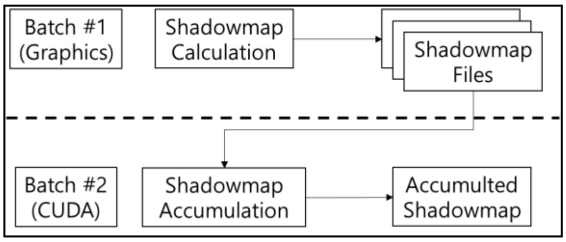
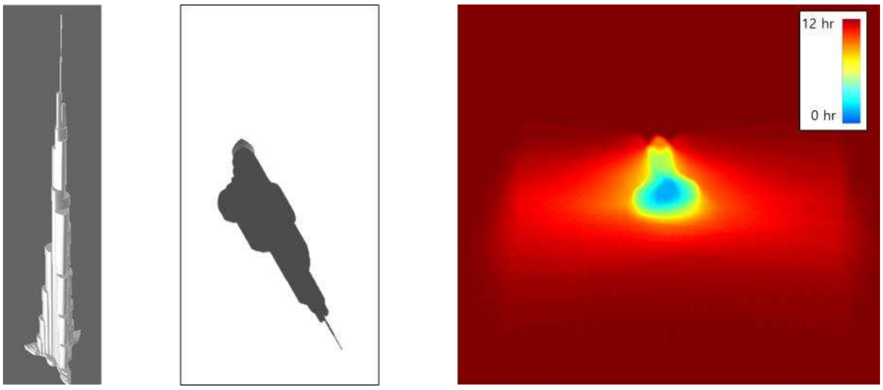


Fig. 1. Summarized Pipeline Architecture of system for Accumulative Sunlight Duration Analyzing

Note that system in [3], [4] can achieve more performance enhancement as the analyzing surface increases and gets complicated. Since the trend of state of the art architecture and building design is getting more complicated for esthetic and environmental purpose, the system can fully fulfill the demand of sunlight analyze on supplicated building as long as 3D model exists. Fig. 2 shows an example of single shadow map acquired at single step, pseudo-colored accumulated shadow map for well know building "Burz Khalifa". It is also showing a 3D model used to construct the map. The 3D model itself contains approximately 24k vertices and 16k faces. Though more sophisticated and high density 3D model can be used, we used example model since it is enough to show characteristics of outline of the building.



(a) 3D Model (b) Single Shadow Map (c) Accumulated Shadow Map

Fig. 2. 3D model of Burz Khalifa, a generated shadow map from single step and pseudo-colored accumulated shadow map for the model

However, though the system can perform analysis on given geometry, it has it's limitations to nature. Since it stores Shadow maps as file at the end of each step, it requires waiting file system to write one shadow map, which is 900KB (1280*720 gray scale image, uncompressed). Though single step analysis file is small to store, at the end, it generates huge amount of data if whole step of analysis is processed. Note that if system requires full year analysis with 15 minute single steps, it must store at least 14k images (approximately 12GB) to file system. With the same manner, it also reads stored file from storage to accumulate whole written shadow maps. Another drawback is that given system performs shadow map calculation with single context, which means it builds shadow map fully sequentially, not in parallel manner. Final limitation is that it utilizes NVidia's CUDA (library) to accumulate shadow maps, however, the library is dependent to hardware.

2 Pipeline refinement

In this section, we introduce new pipeline and system design to increase performance of existing system. We improve performance and give platform independency by implementing accumulator with OpenGL compute shader. It is also to combine separated pipeline to minimize need of file I/O. We also use multiple contexts to fragmentize whole serial steps into separable parallel executable steps. Our refined pipeline is shown in Fig. 3.

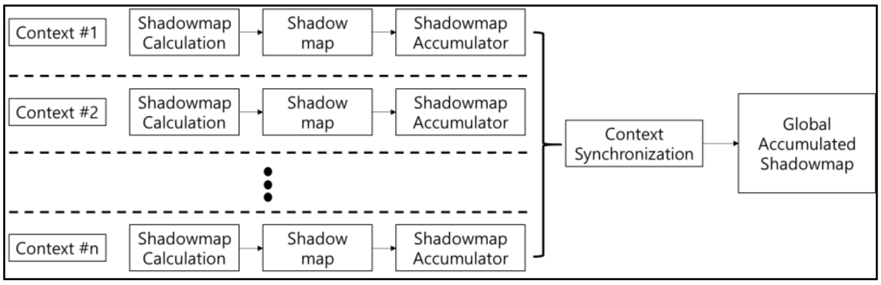


Fig. 3. Refined S for Generating Accumulated Shadow Map

2.1 Adapt to Compute Shader

In the past, main purpose of GPU was to provide graphics pipeline to application for graphics output. However, in these days, GPU is widely used to perform parallel executions by adding more arithmetic operations to chipset. OpenGL compute shader utilizes these architecture and is widely in many purposes. It includes ray tracing computation [5], ambient occlusion computation [6], and additional computation the program requires which can be executed in parallel.

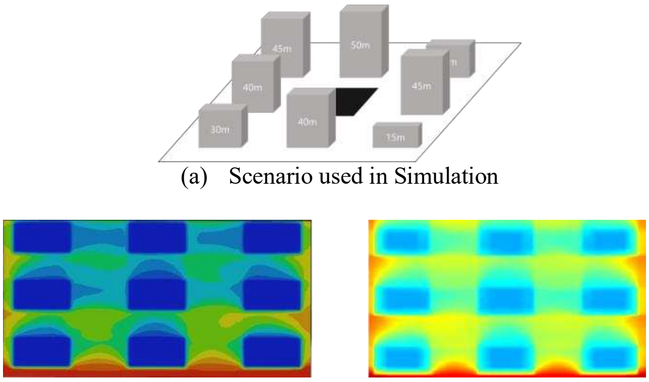
Since the calculated output shadow map is 1280*720 size image and also output to be accumulated shadow map is same type image, we can execute accumulation per pixel, using same computational procedure. When accumulating, the pipeline also stores how many steps have been processed to divide all pixels with that value at the end. By removing unnecessary file I/O, we acquired 2.6times performance enhancement for the same operation mentioned in [4].

2.2 Adapt to Multiple Context

Additionally, we can also divide analysis process into multiple fraction. That is, since effect of sunlight for certain time is independent to another time's effect, analysis for certain period of time can be divided into multiple fractions. After logically dividing time to certain amount, the fraction can be individually processed by multiple threads, and these threading can be implemented by using multiple contexts in graphics application. Multiple context graphics application program can be implemented using GLFW library or can be manually implemented by copying multiple data and executing it in different render pass in Vulkan library. At the end of the execution of each context, a synchronization should be done to generate global accumulated map for whole context.

3 Performance Analysis

To compare performance of our newly designed system, we put DL-Light in base line. DL-light is utility for SketchUp made by De Luminae Lab [7]. DL-Light provides multiple simulations related to sunlight. In this comparison, we used sun exposure calculation module. We gave same geometry (scenario) for all systems, which consists of 9 coalescing different height buildings. Fig.4 shows scenario used in comparison and generated sunlight exposure map for two systems. Table 1 shows execution time of different systems to perform sunlight analysis for the same scenario.



(a) Scenario used in Simulation

(b) result of DL-Light

(c) result of our system

Fig. 4. Scenario used in simulation and results for each system

Table 1. Total execution time of different systems to perform sunlight duration analyze on the same (urban) scenario

System	Execution time
De Luminae	708 seconds
System in [4]	768 seconds
Refined to use Compute Shader	292 seconds
Refined to use Multiple Context	71 seconds

4 Conclusion

In this paper, we refined design of pipeline which analyzes sunlight duration. Refined pipeline design gave up to 10.8 times faster processing time compared to past design, and 9.9 times faster than well-known utility, De Luminae for SketchUp. We also show that our proposed system produce almost same result on calculating sun exposure time compared to commercial software. In the near future, we will optimize batch size of compute shader and adapt to SYCL to support heterogeneous devices and improve performance.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant 2016R1D1A3B03935488 and 2019R1I1A3A01061310).

References

1. Li, D. H.W, Wong, S.L.: Daylighting and energy implications due to shading effects from nearby buildings. Applied Energy 84.12: 1199--1209 (2007)
2. Li, D. H.W., C.C.S. Lau, J.C. Lam.: Predicting Daylight Illuminance by Computer Simulation Techniques. Lighting Research and Technology 36(2): 113--129. (2004)
3. Shin, W., Baek, N.: Design and implementation of a sunshine duration calculation system with massively parallel processing. Advances in Intelligent Systems and Computing 770: 91-97 (2019)
4. Shin, W., Baek, N.: Sunlight Radiation Analysis in Urban Scenario using Layered Accumulative Shadow Map. In: International conference on Culture Technology 2018: 346-349 (2018)
5. Purcell, T., Buck, I., Mark, W., & Hanrahan, P.: Ray tracing on programmable graphics hardware. ACM Transactions on Graphics 21(3): 703-712 (2002)
6. McGuire, M., Osman, B., Bukowski, M., Hennessy, P.: The alchemy screen-space ambient obscurance algorithm. In: ACM SIGGRAPH Symposium on High Performance Graphics. 25-32 (2011)
7. De Luminae Laboratory https://deluminaelab.com

Graph Neural Network with Rejection Mechanism

Bencheng Yan, Chaokun Wang, Gaoyang Guo, Jun Chen
School of Software, Tsinghua University, Beijing 100084, China
Baidu Inc.
{ybc17, ggy16}@mails.tsinghua.edu.cn, chaokun@tsinghua.edu.cn, chenjun22@baidu.com

Abstract. Recently, graph neural networks (GNNs) have achieved great success in dealing with graph-based data. The basic idea of GNNs is iteratively aggregating the information from neighbors, which is actually a special form of Laplacian smoothing. However, most of GNNs fall into the over-smoothing problem, i.e., when the model goes deeper, the learned representations become indistinguishable. It reflects the inability of the current GNNs in exploring the global graph structure. In this paper, we propose a novel graph neural network with rejection mechanism to address this problem. A number of experimental results demonstrate that the proposed model outperforms the state-of-the-art GNN models, and can effectively overcome the over-smoothing problem.

Keywords: GNN, over-smoothing, rejection mechanism

1 Introduction

Graph structure data is ubiquitous in the real world, such as social network [5,7], citation network [2] and graph-based molecules [9]. Recently, graph neural networks (GNNs) have aroused a surge of research interest. The goal of GNNs is to learn representation vectors for nodes in a graph, and then the learned vectors can be used in many graph-based applications, such as link prediction and node classification [7]. The general idea of GNNs is "message propagation", i.e., each node iteratively passes, transforms, and aggregates messages (i.e., features) from its neighbors. Then after k iterations, each node can capture the neighbor nodes' information within k-hops.

However, most of GNN models suffer from the "over-smoothing" problem [4,10]. Specifically, message propagation is proved to be a type of Laplacian smoothing (weighted averaging the features of each node and its neighbors), and stacking too many layers may lead to the representations of nodes indistinguishable and hurt the performance of GNNs [4]. Furthermore, a random walk view to message propagation shows GNNs will converge to a random walk distribution [10], leading to similar conclusions with "over-smoothing".

As a matter of fact, many GNNs are shallow neural networks with only two or three layers, and a limited structure information is learned. Adding additional layers always cannot improve the performance of GNNs, and may even have the opposite effect to GNNs [4,3]. It reflects the inability of the current GNN model in exploring the global graph structure.

In this paper, we focus on dealing with the limitations in current GNNs, i.e., “over-smoothing”, which leads to the inability to explore the global structure. The main contributions of this paper are summarized as follows,

1. We propose a novel neural network, i.e., Graph neural networks with Rejection mechanism (GraphRej), to learn expressive node representations.
2. We design a rejection mechanism which is a simple but effective strategy to avoid over-smoothing of node representations.
3. Extensive experimental results show that the proposed model achieves the state-of-the-art results. Also, the effectiveness of the rejection mechanism in GraphRej is demonstrated.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 proposes GraphRej, our model for graph representation. Section 4 reports the experiments. Section 5 concludes this paper.

2 Related Works

Recently, there are many works focusing on analyzing graph-based problems by leveraging the graph neural networks. GCN [3] is the pilot work, which simplifies the localized spectral filters used in [1]. GraphSGAE [2] takes the representation learning into a formal pattern, i.e., aggregation and combination, and proposes several kinds of aggregation strategies. GAT [9] considers the diversity in neighborhoods, and leverages the self-attention mechanism [8] to effectively select important information in neighborhoods.

Meanwhile, there are several works analyzing the mechanism of the graph neural networks. Li et al. [4] take graph neural networks as a special form of Laplacian smoothing and show the limitation of current GNNs, i.e., GNNs cannot capture the global graph structure due to the over-smoothing problem. Although Li et al. [4] leverage co- and self-training to improve the GNN performance, both co- and self-training are designed to deal with the limited validation issue. Hence, the over-smoothing is still needed to be solved. Xu et al. [10] provide another view of graph neural network. The message propagation in GNNs can be taken as a modified random walk. The distribution of influence score between two nodes will converge to a stationary distribution, which also reflects the over-smoothing problem. Xu et al. [10] propose a dense-connection module to adaptively aggregate neighbors from different hops as well as to avoid the over-smoothing problem.

3 GraphRej

In this section, we first introduce preliminaries. Then the limitation of current GNN is analyzed. At last, we present the proposed model Graph neural network with Rejection mechanism (GraphRej) to overcome this limitation.

3.1 Preliminaries

We represent a graph G as (V, E), where V denotes the node set, and E is the edge set. Let A be the adjacency matrix. The k-hop neighborhood $N_k(v)$ is a set of nodes reaching to node v in k steps exactly. We define $N(v)$ as the 1-hop neighborhood node set. Each node $v \in V$ is associated with a feature vector X_v and a label y_v. The node labels can be taken as supervised information to optimize GNNs. The goal of GNN is to learn a node representation mapping function by using the graph structure and node features.

3.2 Limitation of Current GNNs

Although GNN models outperform many state-of-the-art methods significantly on some benchmarks, there still remain some problems which block the performance of GNNs.

In point of fact, the message propagation scheme in GNNs can be taken as a random walk [10]. The influence score of node u on node v which measures the effect of the input feature of node u to the representations of node v, can be defined as $I(v,u) = \sum_i \sum_j | \partial h_{v,i}^k / \partial h_{u,j}^0 |$ where h_v^k is the representation vector of node v at the k-th layer, and $h_u^0 = X_u$. Then, the expectation distribution of the normalized influence score follows a slightly modified k-step random walk distribution $P(u \mid v,k)$ starting at the root node v [10]. Hence, the message propagates from node u to node v in a random walk pattern.

As we know, the node sequence $(v_t : t = 0,1,..)$ generated by a random walk is a Markov chain. When $k \to \infty$, the probability distribution $p(u \mid v, k \to \infty)$ coverages to a stationary distribution (i.e., $P(o \mid x) \equiv P(o \mid y)$ for any node $x, y, o \in V$) [6]. It means that, for the message propagation scheme, the representation of every node is influenced almost equally by any other node (i.e., $I(x,o) \approx I(y,o)$) when GNNs go deeper (i.e., a large value of k). In other words, the node representation may be over-smoothed, and loses its focus by the information from distant nodes [4], which is called the over-smoothing problem.

Therefore, one of the big limitations of current GNNs is that, most models cannot go deeper (i.e., a deeper version of these model even performs worse than a shallow version), and the best performance of these models is usually achieved within two or three layers [3]. Although a sufficient size of neighborhoods is especially important which allows the model to explore a more complex graph structure and aggregate useful neighbors' information [4,10], current GNNs can only capture limited structure information in a small size of neighborhood. Hence, in this paper, we propose a new neural network architecture to tackle the over-smoothing problem, and capture a more complex graph structure.

3.3 Rejection Mechanism

To address the over-smoothing problem, we propose a simple but effective Rejection Mechanism (RM). As discussed in Section 3.2, the over-smoothing problem is caused by averaging too much information from distant neighbors. Hence, we introduce a learnable hop penalty parameter $c^{(k)} \in R$ (optimized by the backward) at each layer to adaptively control the message flowing from one layer to the next layer (i.e., from one hop to the next hop). This enables the model to reject the messages from distant nodes to avoid over-smoothing.

Firstly, we adopt a message propagation scheme in GraphRej. At the k -th layer, the propagation can be expressed as

$$\begin{aligned} m_v^{(k)} &= Mean(\{h_u^{k-1} : u \in N(v)\}) \\ h_v^{(k)} &= \sigma(W^{(k)} Combine(m_v^{(k)}, h_v^{(k-1)})) \end{aligned} \tag{1}$$

where $h_v^{(k)}$ is the representation (i.e., "message") of node v at the k -th layer and $m_v^{(k)}$ can be taken as the integrated message representation from neighbors. $W^{(k)}$ is the weight matrix, and σ is a nonlinear function. The initialization message $h_v^0 = X_v$, $Mean$ is a message propagation function which averages information from neighbors, and $Combine$ is a combination function that combines the information from different hops.

Then, the rejection mechanism focuses on the combination function. Specifically, at the k -th layer, the hop penalty parameter $c^{(k)}$ is applied into the integrated message representation from neighbors, then the combination function can be rewritten as

$$\begin{aligned} c^{(k)'} &= sigmoid(c^{(k)}) \\ h_v^{(k)} &= \sigma(W^{(k)}(h_v^{(k-1)} \oplus c^{(k)'} \otimes m_v^{(k)})) \end{aligned} \tag{2}$$

where $W^{(k)}$ is a weight matrix, $c^{(k)'} \in [0,1]$ is the rescaled penalty parameter, and $\otimes$ refers that each element of the vector $m^{(k)}$ multiplies the rescaled hop penalty parameter $c^{(k)'}$, $\oplus$ refers to element-wise addition. Intuitively, the value of $c^{(k)'}$ can be regarded as a gate to influence the message propagation from neighbors to the center node.

Analysis. To fully understand the rejection mechanism, we provide an example (shown in Figure 1) to illustrate how it works for avoiding the over-smoothing problem. For simplicity, we take a chain graph containing four nodes as an example, and only consider the predecessor nodes as 1-hop neighbors. Considering a three-layer GNN with the mean aggregation function and the proposed combination function with RM, the propagation in the k -th layer for this chain graph can be expressed as (ignoring the weight matrix and non-linear function in Equation 2)

$$\begin{aligned} m_i^{(k)} &= h_{i-1}^{(k-1)} \\ h_i^{(k)} &= h_i^{(k-1)} \oplus c^{(k)} m_i^{(k)} \\ &= h_i^{(k-1)} \oplus c^{(k)} h_{i-1}^{(k-1)} \end{aligned} \tag{3}$$

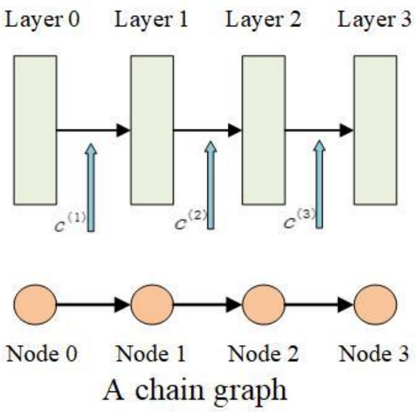


Fig. 1. An example of the reject mechanism in a chain graph.

where $i, k \in \{1,2,3\}$. For expressing convenience, we ignore the symbol $\otimes$ in Equation 2 and $c^{(k)}$ refers to the rescaled parameter $c^{(k)'}$.

Then the final representation of Node 3, obtained from the last layer, can be represented as

$$\begin{aligned} h_3^{(3)} &= h_3^{(0)} \oplus (c^{(1)} + c^{(2)} + c^{(3)}) h_2^{(0)} \\ &\oplus (c^{(1)}c^{(2)} + c^{(1)}c^{(3)} + c^{(2)}c^{(3)}) h_1^{(0)} \\ &\oplus c^{(1)}c^{(2)}c^{(3)} h_0^{(0)} \end{aligned} \tag{4}$$

It reveals that the influence of the distant Node 0 to the representation of Node 3 is punished by $\prod_{k=1}^{3} c^{(k)}$, and $c^{(k)} \in [0,1]$. Multiple multiplications will lead to a small value, which adaptively controls the messages flowing from distant nodes (e.g., Node 0) to the representation nodes (e.g., Node 3).

Hence, the benefits of the proposed rejection mechanism can be summarized as follows. (1) When stacking layers to build a deep GNN, the information from distant nodes will be more likely to be punished or even rejected. It is a simple but effective way to avoid the over-smoothing problem. (2) The information from distant nodes can also affect the node representation, which helps to capture global structure. (3) The combination of the hop penalty parameters leads to adaptively aggregate information from different hops, which contributes to build a more powerful deep graph model.

Finally, at the last layer, a cross-entropy loss is adopted to optimize the parameters of the proposed model.

4 Experiments

We evaluate the benefits of GraphRej against a number of state-of-the-art graph neural networks with the goal of answering the following questions:

Q1 How does the GraphRej perform in comparison to the state-of-the-art GNNs?

Q2 Can the proposed rejection mechanism avoid the over-smoothing problem indeed?

Table 1. Data Sets Information.

Data Sets	#Nodes	#Edges	#Class	#Features
Cora	2,708	5,429	7	1,433
Citeseer	3,327	4,732	6	3,703
Pubmed	19,717	44,338	3	500

Table 2. Node Classification Accuracy (%) for two-layers based GNNs

Method	Cora	Citeseer	Pubmed
GraphSGAE	76.20	71.00	84.54
GCN	79.06	71.90	80.70
GAT	76.66	70.85	84.62
JK	74.82	71.53	84.75
GraphRej	**80.44**	**71.97**	**85.23**

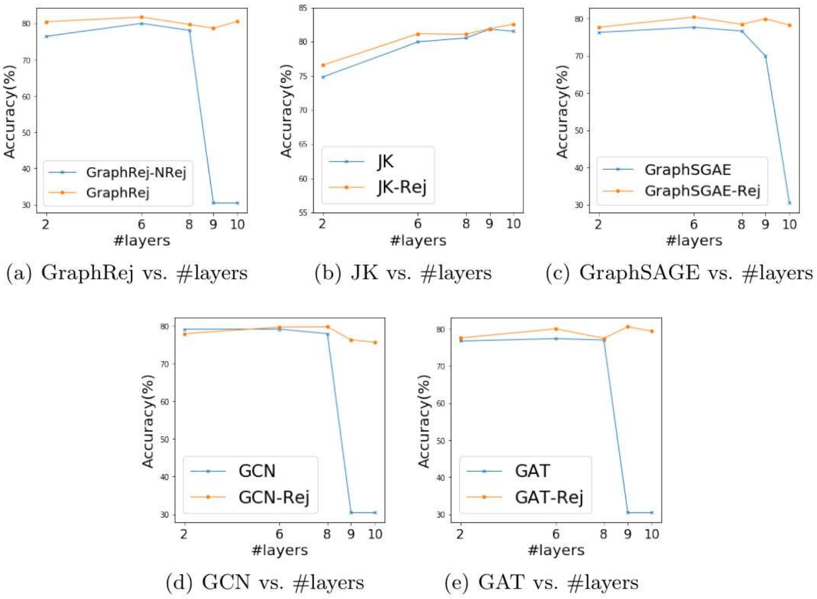
4.1 Data Sets

To adequately evaluate the performance of our model and baselines, we apply them on three real-world data sets. i.e, Cora, Citeseer, and Pubmed. All of them are citation graphs. Each node represents a paper associated with an attribute vector indicating the statistic information of words in this paper. The statistics of the data sets are summarized in Table 1.

4.2 Baselines

In the performance comparison, we consider the state-of-the-art baselines based upon GNNs. **GCN** [3] is a graph convolution neural network, which simplifies the localized spectral filters used in [1], and extracts the 1-localized information for each node in each convolution layer. **GraphSAGE** [2] extends GCN to a more general way, and introduces two basic functions, i.e., the aggregation function and the combination function. **GAT** [9] employs the idea of self-attention [8] to filter important messages from neighbors. **JK** [10] is a research work to deal with the over-smoothing problem, which uses a dense-connection module to adaptively aggregate neighbors from different hops.

Setting. We set the dimension of representation vector to 16 for all models. We use the pooling aggregation in GraphSAGE, which achieves the best performance compared to other aggregation strategy (as discussed in [2]). We use the Maxpool for JK for the same reason as GraphSAGE (i.e., high and adaptive performance discussed in [10]). All models are trained using the Adam SGD optimizer with an initial learning rate of 0.01. We use dropout rate $d = 0.2$ for all layers. For all data sets, we split 20% nodes as the training nodes and 40% nodes as the validation nodes, and the rest 40% as the test nodes. We use an early stopping strategy on both the model loss and accuracy score on the validation nodes, with a patience of 20 epochs.



(a) GraphRej vs. #layers (b) JK vs. #layers (c) GraphSAGE vs. #layers

(d) GCN vs. #layers (e) GAT vs. #layers

Fig. 2. The results of GNNs combing with the rejection mechanism. "XX-Rejrefers to the model XX with the rejection mechanism, and GraphRej-NRej refers to the version of GraphRej without the rejection mechanism.

4.3 Results for Node Classification Task

To fairly evaluate the performance of GraphRej and baselines on the node classification task, we report the classification accuracy (shown in Table 2) of these models with the same two layers (the performance with different layers is analyzed in Section 4.4). These results provide positive evidence to **Q1**: GraphRej consistently performs significantly better than previous models on all data sets. Specifically, GraphRej achieves an improvement ranging from 0.07% to 5.62% over all data sets. It shows the effectiveness of GraphRej to learn a more meaningful node representation.

4.4 Evaluation of Rejection Mechanism

Actually, the proposed rejection mechanism (RM) is a general idea to avoid over-smoothing. Hence, to evaluate the benefits of the rejection mechanism (answering the question **Q2**), we apply it to other GNN models, and report the classification accuracy of different models with or without the rejection mechanism when stacking more and more layers. Additionally, we provide the results of GraphRej with or without the rejection mechanism. The results on Cora in terms of classification accuracy are presented in Figure 2 (similar performance trends are also observed on other data sets).

We observe that RM significantly improves the performance of all GNN models. Specifically, (1) when we gradually increase the number of layers, most of original models fail with the over-smoothing problem, i.e., when it comes to a deeper model (e.g., the layer of GNNs is 9 or 10 in Figure 2), the performance of most of these models sharply decrease, which also supports the discussion mentioned in Section 3.2. When we apply RM to these failed models, all of these models obviously overcome the over-smoothing problem, and achieve high performance in a deep version. Thus, it indicates the rejection mechanism is indeed beneficial to overcome the over-smoothing problem. (2) With the same number of layers, all models outperform the corresponding models without RM. Especially, although JK [10] is also designed to avoid the over-smoothing problem, combining RM still can improve the performance of JK. It demonstrates the effectiveness of RM.

5 Conclusion

In this paper, we first address the limitations of current GNNs, i.e., the need and the bottleneck both introduced by the global information. Then a rejection mechanism is designed, which is a simple but effective way to avoid the over-smoothing problem. Exhaustive results demonstrate that the proposed model achieves the state-of-the-art results. In the future, we will put more effort to analyze how to design a good architecture for GNNs with RM, to capture graph information as much as possible.

6 Acknowledges

This work is supported in part by the National Natural Science Foundation of China (No. 61872207) and Baidu Inc. Chaokun Wang is the corresponding author.

References

1. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural graphs with fast localized spectral filtering. NIPS pp. 3844–3852 (2016)
2. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: NIPS. pp. 1024–1034 (2017)
3. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
4. Li, Q., Han, Z., Wu, X.M.: Deeper insights into graph convolutional networks for semi-supervised learning. In: AAAI (2018)
5. Lou, Y., Wang, C.: OSMAC: Optimizing Subgraph Matching Algorithms with Community Structure.. In: ICDE. pp. 1750–1753. (2019)
6. Lovasz, L.: Random walks on graphs: A survey. Combinatorica (2001)
7. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: KDD. pp. 701–710. ACM (2014)
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
9. Veliˇ ckovi´ c, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph Attention Networks. ICLR (2018), https://openreview.net/forum?id=rJXMpikCZ
10. Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.i., Jegelka, S.: Representation learning on graphs with jumping knowledge networks. In: ICML. pp.5449–5458 (2018)

Constructing the RB+-tree on Large Scale Multi-dimensional Data with Hadoop-MapReduce

Huynh Cong Viet Ngu[1] , Keon Myung Lee[1]

[1] Department of Computer Science, Chungbuk National University, Cheongju, South Korea
{huynhcongvietngu1.1@gmail.com, kmlee@cbnu.ac.kr}

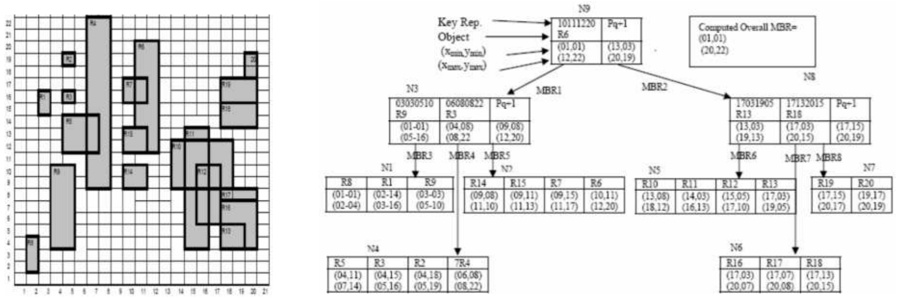
Abstract. By combining the strengths of B+-tree and R-tree, the RB+-tree is gradually considered as a replacement candidate for R-tree in indexing and querying to multi-dimentional data. However, it also usually takes much time to build the RB+-tree for a large amount of the spatial data that is being generated significantly from many various devices, such as smartphone, satellites, and medical devices. With the advent of Hadoop-MapReduce, a parallel and distributed programming model, it motivated our efforts to improve the RB+-tree construction time. In this paper, we propose a parallel RB+-tree construction schema based on a Hadoop framework. The proposed schema first divides the whole data set into partitions so that each partition have nearly equal amount of data, then in the second stage, from the partitions that are divided in the previous stage, the local RB+-trees are built in parallel, and combine them into the final RB+-tree that covers whole dataset in the last stage. Therefore, our schema gives a very efficient spatial index structure while reducing the construction time of the RB+-tree significantly

Keywords: RB+-tree, Hadoop, Multi-dimensional indexing, MapReduce.

1 Introduction

In recent years, beside traditional data, a large amount of the spatial data that is being generated significantly from many various devices, such as smartphone, satellites, and medical devices. In order to manage this amount of data efficiently, most of the spatial database management system (SDBMS) use R-tree or its variants as a mechanism for indexing and querying data quickly from its spatial position in the database. However, the R-tree has some drawbacks such as the page-split operation is very expensive and in search opeartion, from root-node to leaf-node (data), it may follow several paths. From these reasons, by combining the strengths of B+-tree [1] and R-tree [2], the RB+-tree [3] is gradually considered as a replacement candidate for R-tree in indexing and querying to multi-dimentional data. Similar to R-tree, in RB+-tree, each node contains a fixed number of entries (records), each of which consists of a Minimum bounding Rectangle (MBR) and a pointer to child node if it's a non-leaf node or data if it's a leaf node. On the other hand, to improve the search performance, the structure of the RB+-tree is designed so that the number of node that need to be traversed is the least.

In RB+-tree, the structure of the leaf-node and non-leaf node are different. In non-leaf node structure, let $\mathbf{M_{nl}}$ be the maximum of entries in node, all nodes have at least $\mathbf{M_{nl}/2}$ children, except the root-node has at least **2** or $\mathbf{M_{nl}/2}$. In leaf-nodes, let $\mathbf{M_l}$ be the maximum of entries (data), then all nodes have at least $\mathbf{M_l/2}$**.** The example of the RB+-tree for the 2D spatial objects is shown in Fig.1 [1] as below



(a) **(b)**

Fig. 1. **(a)** shows twenty sample data in 2D-space while **(b)** shows RB+-tree structure of the spatial objects shown in **(a)**.

In fact, in applications where a large amount of data is available, it take a lot of time for the RB+- construction since the data is inserted sequentially. Since its release in April 2005, Hadoop-MapReduce [4] was adopted as an optimal solution for massive datasets processing, therefore, it motivated our schema for RB+-tree construction that take advantage of the parallelism of the MapReduce model to reduce the construction time while the quality of the RB+-tree is also efficient. The rest of the paper is organized as follow. The section II describes our schema in detail and the last section is our conclusion and future work.

2 Proposed schema for Parallel RB+-tree Construction

The main considerations in our schema are reduce the construction time of the RB+-tree and minimize the area of the MBRs of the non-leaf nodes that are not covered by MBRs. Our schema consists of three stages which are performed in a sequential manner. First, it divides the whole data set into smaller partitions so that each partition have equal amount of data, then in the second stage, from the partitions that are divided in the previous stage, the local RB+-trees are built in parallel. The first two stages are implemented in MapReduce model. And in the last stage, we combine all of local RB+-trees that are built in the second stage into the final RB+-tree that covers whole dataset. Therefore, our schema gives a very efficient spatial index structure while reducing the construction time of the RB+-tree significantly. Our schema is shown in the **Fig. 2.**

Note that in the dataset, all objects lie in 2D-space, each object is represented by <**o.k**, **o.v**>, where **o.v** is the position of the object that is represented by (x_{min}, y_{min}, x_{max}, y_{max}) and **o.k** is the object's identifier.

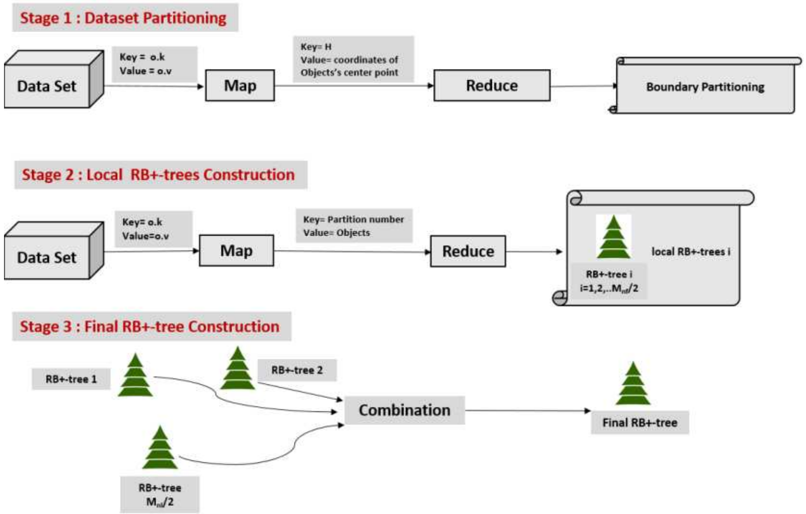
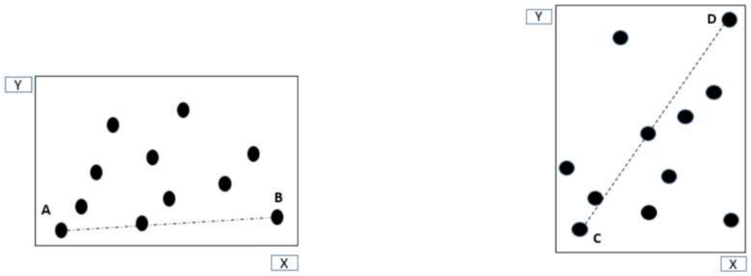


Fig. 2. Proposed schame for Parallel RB+-tree construction in Hadoop-MapReduce

2.1 Dataset Partitioning

This stage is implemented in MapReduce model. The main goal in this stage is determine a list of boundary in 2D-space so that the number of objects on each partition are equal. In addition, in order to minimize the area of the overall MBR that cover all objects on each partition, first we determine coordinate has the two most centers between the objects, then data is partitioned based on that coordinate, as shown in Fig.3. Note that the number of partition is the number of entries in non-leaf node of the RB+-tree as we defined before **M$_{nl}$/2**.



(a)

(b)

Fig. 3. (a) Data is partitioned by X-coordinate **(b)** Data is partitioned by X-coordinate.

Instead of reading all data from input file, a sampling technique is applied. In this stage, read whole dataset will take very long time while the accuracy in boundary determination is almost no difference.

Firstly, the Mappers take sample objects from the input file at the given ratio, then in each Mapper, center point's coordinates of every objects are calculated. In order to send all outputs from Mappers to single Reduce, the intermediate key of Mappers is the constant **H**.

In Reducer, the list of center points of the objects are sorted in ascending as shown in Fig.3, and determine the coordinate has the two most centers between the objects. In Fig.3-a, since the distance between A and B is largest, x-coordinate is used for data partitioning, in Fig.3-b is the opposite case. Then based on that coordinate, a list of splitting point $\mathbf{T = M_{nl}/2 - 1}$ is determined as the output of this stage, as shown in Fig.4.

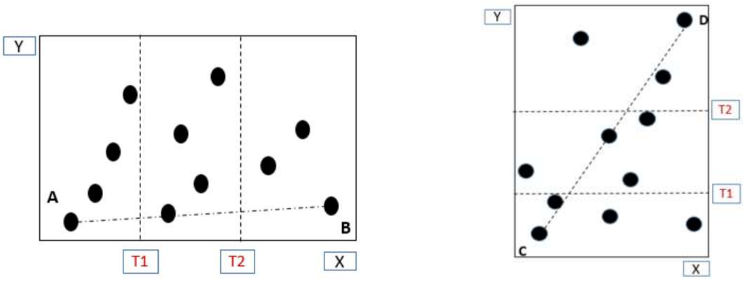


Fig. 4. A list of splitting point **T** is determined based on the chosen coordinate

2.2 Local RB+-trees Construction

In this stage, the main goal is the individual RB+-trees are built simultaneously from the whole data set. Thus, it's the most intensive computational processing stage in our schema, so it will take much time to complete.

In the Map phase, Mappers read all data, for each object, they calculate the center point, then partition it in one of $\mathbf{M_{nl}/2}$ groups based on a list of splitting point **T** that is determined in previous stage. After that, every partition will be executed by a different Reducer. In each Reducer, a local RB+-tree is constructed independently. The output of every Reducer is the root-node of constructed RB+-tree. The MapReduce function of this stage is shown in table 1

Table 1. MapReduce function for Parallel Local RB+-tree Construction

Function	Input data	Output(Key,Value)
Mappers	(o.k, o.v)	(partition Number, Objects)
Reducers	(partition Number, list(Objects))	Local RB+-tree, root

2.2 Final RB+-tree Construction

In this stage, a final RB+-tree will be consolidated from local RB+-trees that were built in the previous stage under a single root. This phase is executed outside the cluster because it's fairly simple as shown in Fig.2.

3 Conclusion and Future Work

In this paper, the proposed schema that has three stages for parallel RB+-tree construction, in which, the first two stages are implemented in MapReduce model, while the last stage is executed outside the model because it does not require the high computational. With this work, we hope contribute to the improvement of the storage and query of the spatial data in the context of the data size is still increasing dramatically.

In RB+-tree, the insertion operation is implemented with top-down approach where the objects are inserted sequentially as ther arrived. However, in fact that there are many application where data is available, it should be constructed with bottom-up approach that can minimize the area of MBR of the non-leaf nodes. For this reason, in near future, we will develop an appropriate packing technique for RB+-tree that can improve the quality of the RB+-tree. Morever, we will continue improve our schematic as well as optimize the storage and query performance for spatial data in particular and all types of data in general, in the context of their size is still increasing dramatically in the Industry 4.0 era.

Acknowledgment

This work was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea, Republic of Korea (Grant no.: NRF-2017M3C4A7069432).

References

1. Douglas, C.: The Ubiquitous B-Tree. Journal of ACM Computing Surveys (CSUR). 121-137 (1979)
2. A Guttman: R-trees-A Dynamic Index Structure for Spatial Searching. ACM SIGMOD. 47-57 (1984)
3. M.Fekihal, I.Jaluata, I.Osman: RB+-Tree: A Multidimensional Index Structure. In: 5th IEEE International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia (2009)
4. T White: Hadoop: The definitive guide. O'Reilly Media Inc, USA (2012)

An Autocalibration Model of Low-Cost Sensors for Higher Accuracy of Fine Particle Data Analysis

Ah Young Kang, Hye Jin Lim, Yeong Hyeon Gu, Seong Joon Yo[1]

Department of Computer Engineering,
Sejong University,
05006 209, Neungdong-ro, Gwangjin-gu, Seoul Korea
sjyoo@sejong.ac.kr

Abstract. Recently, the Korean people have become concerned about the more frequent occurrence of highly concentrated fine particles and there is a trend to measure fine particles using low-cost portable sensors. But measurements by such low-cost sensors are considerably less accurate than measurements by AirKorea's sensors. Against this backdrop, this study suggests models to automatically calibrate measurements by low-cost sensors against measurements by AirKorea's sensors for the purpose of improving the accuracy of low-cost sensors. Low-cost sensors were installed at a platform and a waiting room of an underground subway station (Suyu Station, Seoul subway line 4) to measure fine particle concentrations on an hourly basis. The measurements were used to design and develop models that automatically calibrate themselves against measurements by AirKorea's sensors. The results of the experiment are as follows. In the case of the subway platform, the Bagging model showed the highest performance with a mean absolute percentage error (MAPE) of 15.30 % and a root-mean-square deviation (RMSE) of 35.88 %, while in the case of the waiting room, the Gradient Boosting (GB) model showed the highest performance with a MAPE of 15.30 % and a RMSE of 25.90 %. Consequently, the autocalibration models suggested in this study are expected to contribute to the provision of more accurate measurements of fine particles in daily life.

Keywords: fine particle, autocalibration, machine learning, AirKorea, low-cost sensor

1 Introduction

Recently, the Korean people have become concerned about the more frequent occurrence of highly concentrated fine particles [1]. Fine particles are airborne pollutants that have an aerodynamic diameter of 10㎛ or less. When inhaled, they impair respiratory and cardiovascular systems. As the problem of fine particles becomes more serious, the government has recognized the problem as one that could have serious social repercussions and which needs to be urgently addressed. Accordingly, the Special Act on Fine Particle Reduction and Management was enacted [2]. As it is impossible in actuality to entirely eliminate fine articles, it is

[1] Corresponding Author

important to prevent harm and respond to this problem of fine particles through accurate forecasts.

Recognizing the importance of forecasts of fine particulate matter, this study developed models that automatically calibrate measurements by low-cost sensors against measurements by AirKorea's sensors to obtain values close to those by AirKorea's sensors. Two low-cost sensors were installed at Suyu Station, Seoul subway line 4, one at the platform and the other at the waiting room, in order to obtain air quality information from April 25 to May 25, 2019. The obtained air quality information and PM10 values measured by AirKorea's sensors were used as data in the experiment. The air quality information included data such as PM10, PM2.5, CO2, HCHO, VOC, temperature, and humidity. And this study developed autocalibration models using the following algorithms: Linear Regression (LR), RandomForest (RF), ExtraTrees Regressor (ETR), AdaBoost (AB), Bagging, Gradient Boosting (GB), and Multi-Layer Perceptron (MLP). When the performance of the models were compared, in the case of the platform, the Bagging model showed the highest performance with a mean absolute percentage error (MAPE) of 15.30 % and a root-mean-square deviation (RMSE) of 35.88 %, while in the case of the waiting room, the Gradient Boosting (GB) model showed the highest performance with a MAPE of 15.30 % and a RMSE of 25.90 %.

In this study, Chapters 2 to 6 describe the following: studies related to fine particles (Chapter 2); datasets used to develop autocalibration models (Chapter 3); algorithms used to develop autocalibration models (Chapter 4); the results of the study (Chapter 5); and the conclusion of the study and a further study plan (Chapter 6).

2 Related Studies

A number of studies have been conducted on fine particle concentrations in Korea and other countries. In [3], environment data (SO2, CO, NO, NO2) and weather data (temperature, relative humidity, wind speed) were used to forecast an average fine particle concentration on a daily basis, and the performance of five linear models was compared and analyzed. In [4], fine particle figures in Mediterranean countries were forecasted using data sets that consist of hourly fine and ultrafine particle levels, fine particle warnings and weather data, as well as artificial neural networks (ANN). In [5], concentrations of fine and ultrafine particles were analyzed using deep neural networks and K-means Clustering. In [6], a fine particle prediction model was suggested using weather datasets, linear regression analysis, artificial neural analysis and a long-term short-term memory (LSTM) network. In [7], data gained by measuring airborne pollutants every 15 minutes were formed into datasets, and then a fine particle prediction model was suggested using Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), Vector Autoregressive Moving Average (VARMA), and AutoRegressive Integrated Moving Average (ARIMA). In [8], an analysis of fine particles was performed by integrating weather data, vehicle speed and traffic volume as variables, and a fine particle prediction model was suggested using a genetic algorithm. In [9], a fine particle prediction model was suggested by integrating airborne pollutants and weather data as variables and using a Gradient Boosting

Machine (GBM) algorithm.

In previous studies, machine learning and deep learning models were used to predict fine particle concentrations. In this study, measurements of low-cost sensors have considerably lower precision than those of AirKorea's sensors and therefore, the measurements by low-cost sensors have limited reliability. As there is a growing distribution of low-cost devices measuring fine particles, improvement in their measurement accuracy is required. Against this backdrop, this study suggests models to automatically calibrate measurements by low-cost sensors against measurements by AirKorea's sensors.

3 Dataset

This chapter describes the data used to develop autocalibration models as well as input variables used additionally. Table 1 shows input variables used to develop a model.

Table 1. Variables

Category	Feature
Low-cost sensor variable	PM10 [2]
	PM2.5 [3]
	CO2 [4]
	HCHO [5]
	VOC [6]
	TEMPERATURE
	HUMIDITY
Derived variable	Clustering Value
Target variable	Air-Korea's PM10

3.1 Data Collection and Processing

This study performed an experiment in which measurements by low-cost fine particle sensors were automatically calibrated. The measurement location was Suyu Station, Seoul subway line 4. Data were collected from hourly measurements by two low-cost sensors, one installed at the platform and the other at the waiting room, and measurements by AirKorea's sensors were recorded at the same spots, from April 25 to May 25, 2019. When the PM10 value was 0 or there was no PM10 value in the low-cost sensor variables or when AirKorea's PM10, the target variable, was 0 or had no value, the data were removed and then clustering was performed to add variables.

[2] PM10 is particulate matter less than 10 ㎛ in diameter and called fine particles.

[3] PM2.5 is particulate matter less than 2.5 ㎛ in diameter and called ultrafine particles.

[4] Carbon dioxide

[5] Formaldehyde

[6] Volatile Organic Compounds (VOCs)

A total of 693 cases of data measured at the platform and a total of 731 cases of data measured at the waiting room were used. Low-cost sensor variables include PM10, PM2.5, CO2, HCHO, VOC, and temperature and humidity values measured by the low-cost sensors. And the clustering value was used as a derived variable, and AirKorea's PM10 value was used as a target variable.

3.2 Adding Variables Using Clustering

In this study, features obtained from clustering were used as additional variables to calibrate measurements by low-cost measurement sensors and make these measurements more closely resemble the measurements by AirKorea's sensors. Clustering is a statistical technique that measures similarity of objects and groups a set of objects in such a way that objects in the same groups are more similar to each other than to those in other groups, while defining the similarity between objects in the same group and the difference between objects belonging to different groups. Low-cost sensor variables at the platform and the waiting room were clustered into five groups, which were then added to input variables.

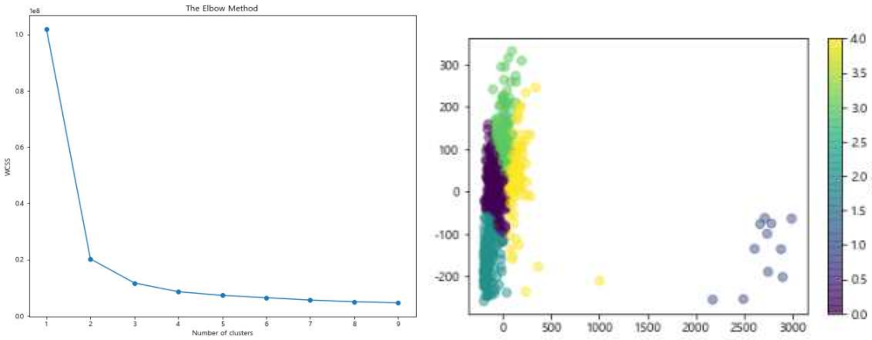


Fig. 1. Results of clustering (data measured at the platform)

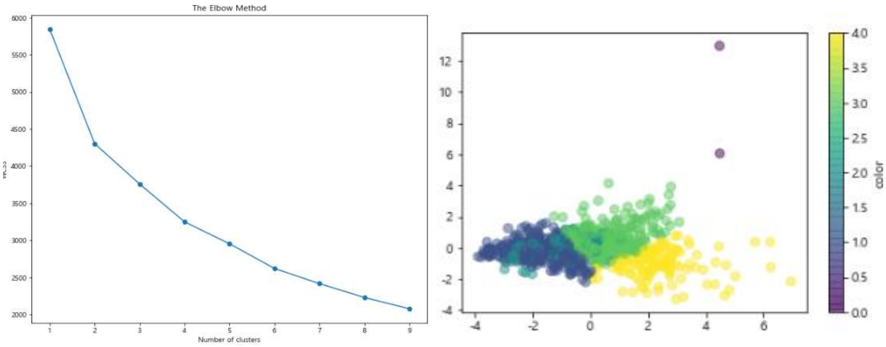


Fig. 2. Results of clustering (data measured at the waiting room)

4 Autocalibration Models

This chapter explains the algorithms used to develop autocalibration models.
Linear regression is a statistical technique used to evaluate or predict correlations between variables. It is divided into two categories: simple linear regression with a single independent variable and multiple linear regression with multiple independent variables. Random Forest is an ensemble method developed for classification or regression that reduces the chance of overfitting the data. It randomly extracts samples from given data to generate multiple decision trees, which are used to predict results through such processes as voting and standardization. The ExtraTree algorithm is a machine learning algorithm, similar to RandomForest. The ExtraTree algorithm splits nodes randomly using possible features to further randomize tree building and then selects the optimal node. Bagging Regressor is a machine learning method that reduces model dispersion. This method is used to shorten the variance in the case of a large variability. AdaBoost Regressor is a learning machine method, and the output of the other learning algorithms ("weak learners") is combined into a weighted sum that represents the final output of the boosted classifier. MLP Neural Network is a deep learning model developed to enable learning of data that is not linearly separable, through multiple layers between the input and output layers.

5 Experiment

Data collected from April 25 to May 24 were used as the whole learning data, and the 24-hour data collected on May 25 were used as test data. The accuracy of automatically calibrated values against actual measurements was determined based on absolute percentage errors (MAPEs) and root-mean-square deviations (RMSEs).
Tables 2 and 3 shows the accuracy of autocalibration models developed using the seven algorithms in relation to data collected at the platform and the waiting room, respectively. In the case of the platform, the Bagging model showed the highest performance with a mean absolute percentage error (MAPE) of 15.30 % and a root-mean-square deviation (RMSE) of 35.88 %, while in the case of the waiting room, the Gradient Boosting (GB) model showed the highest performance with a MAPE of 15.30 % and a RMSE of 25.90 %. In Figs. 3 and 4, the graphs show that values obtained through autocalibration models are closer to measurements by AirKorea's sensors than measurements of low-cost sensor models.

Table 2. Results of the Autocalibration Models (at the platform)

Model	MAPE	RMSE
Linear regression	31.61	53.47

RandomForest	18.40	39.7
GradientBoostingRegressor	19.37	43.37
Extratree	19.80	39.87
BaggingRegressor	**15.30**	**35.88**
AdaBoostRegressor	17.67	40.02
Multilayer perceptron	35.47	59.90

Table 3. Results of the Autocalibration Models (at the waiting room)

Model	MAPE	RMSE
Linear regression	18.25	22.75
RandomForest	17.55	29.24
GradientBoostingRegressor	**15.21**	**25.90**
Extratree	18.91	23.74
BaggingRegressor	16.37	25.84
AdaBoostRegressor	18.61	27.92
Multilayer perceptron	19.36	22.38

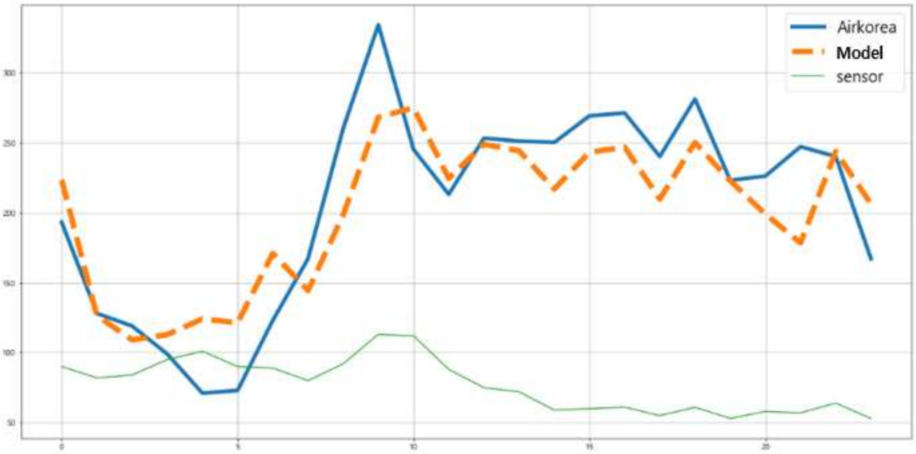


Fig. 3. Comparison of Measurements by Low-Cost Sensors Before and After Calibration against Measurements Provided by AirKorea (at the platform)

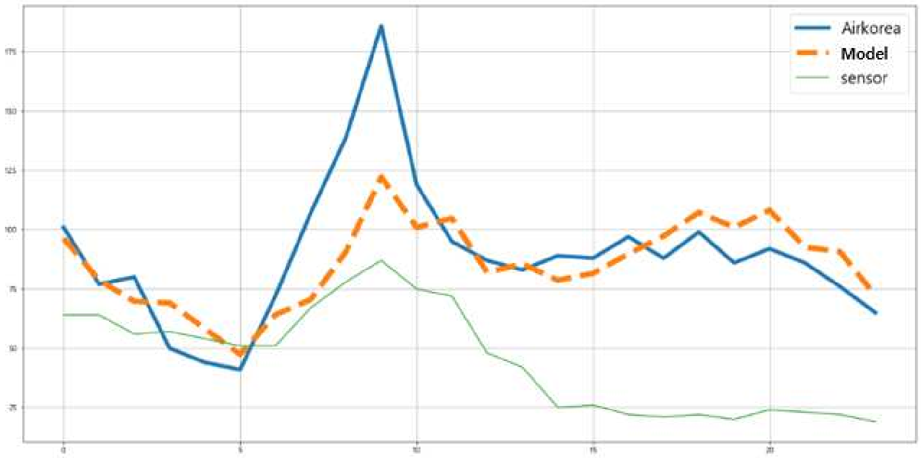


Fig. 4. Comparison of Measurements by Low-Cost Sensors Before and After Calibration against Measurements by Provided by AirKorea (at the platform)

6 Conclusion and a Future Work

This study suggested models that automatically calibrate measurements by low-cost sensors against PM10 measurements provided by AirKorea, using the data collected for a month through low-cost sensors installed at the platform and the waiting room of a subway station. The used algorithms include the algorithm of Linear Regression (LR)—a statistical technique; the algorithms of RandomForest (RF), ExtraTrees Regressor (ETR), AdaBoost (AB), Bagging, Gradient Boosting (GB)—machine learning techniques; and the algorithm of Multi-Layer Perceptron (MLP)—a deep learning technique. The results of the experiment show that the Bagging model had the highest performance in the case of the platform, while the Gradient Boosting (GB) had the highest performance in the case of the waiting room.

It can be seen that the use of autocalibration models for fine particles suggested by this study will decrease the gap between measurements by low-cost sensors and measurements by AirKorea's sensors, thus increasing the accuracy of measurements by low-cost sensors. Because autocalibration was possible when only a month of data were used, the suggested autocalibration models are expected to contribute to more accurate forecasts if applied to low-cost sensors. There were insufficient datasets, which placed a limitation on the application of a deep learning model. In a further study, additional data will be secured and a long-term short-term memory (LSTM) algorithm, which shows high performance with time series data, will be applied for greater accuracy.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT)
(2019-0-00136, Development of AI-Convergence Technologies for Smart City Industry Productivity Innovation)

References

1. Min Ju Yeo, Yong Pyo Kim: Trends of the PM10 Concentrations and High PM10 Concentration Cases in Korea. vol. 35, pp. 249-264. Korean Society for Atmospheric Environment (2019)
2. www.law.go.kr/법령/미세먼지저감및관리에관한특별법/(15718,20180814)
3. J. C. M. Pires, F. G. Martins.: Prediction of the Daily Mean PM10 Concentrations Using Linear Models. vol. 4, pp. 445-453. American Journal of Environmental Sciences (2008)
4. Gianluigide Gennaro.: Neural network model for the prediction of PM10 daily concentrations in two sites in the Western Mediterranean. vol. 463–464, pp. 875-883. Science of The Total Environment (2013)
5. M.A.Elangasinghe.: Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. vol. 94, pp. 106-116. Atmospheric Environment (2014)
6. R. O. Sinnott and Z. Guan.: Prediction of Air Pollution through Machine Learning Approaches on the Cloud. 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), Zurich. pp. 51-60 (2018).
7. P.J. García Nieto, F. Sánchez Lasheras, E. García-Gonzalo, F.J. de Cos Juez.: PM10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study. vol. 621, pp. 753-761. Science of The Total Enviroment. (2018)
8. Uroš Lešnik, Domen Mongus, David Jesenko,: Predictive analytics of PM10 concentration levels using detailed traffic data. vol. 67, pp. 131-141. Transportation Research Part D: Transport and Environment (2019)
9. Georgia Miskell, Woodrow Pattinson, Lena Weissert, David Williams.: Forecasting short-term peak concentrations from a network of air quality instruments measuring PM2.5 using boosted gradient machine models. vol. 242, pp. 56-64. Journal of Environmental Management (2019)

An Effective Scheme For Automatic Adjustment and Update Traffic Signal

Khoa Thi-Minh Tran[1] , Dinh-Phuc Pham[2]

[1,2] Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam
[1] ttmkhoa@iuh.edu.vn, [2] phamdinhphuc90@gmail.com

Abstract. This paper presents a scheme, named S-W-H scheme, for automatic adjustment and update the traffic signal, including state and duration time at the interections in a traffic system. The aims of this scheme is to provide flexible, and low-maintainance wireless solution for obtaining traffic-related data that can be used for determining traffic signal states, calculating traffic signal duration time. This scheme is a combination of support vector machine, webster calculation, and hertistic intersection selection method to handle collected traffic data from sensor nodes. The system with S-W-H applied scheme has been tested and simulated under Matlab simulator.

Keywords: Wireless Sensor Network (WSN), Intelligent Transport System (ITS), Support Vector Machine, Webster, HISM

1 Introduction

In Vietnam, transportation is one of the most issue that interest to many social sectors. According to statistic report of Vietnamese Traffic Police Department, in 2015, there were more than 22,827 traffic accidents that killing 8,727 people and injuring 21,069 people. Including 22,326 cases of roads accidents accounted over 75% of the accidents of all the country. A lot of traffic jams lasted long, mainly due to traffic accidents (44%), traffic congestion (24,4%), and other causes (31,7%). Hence, itelligent traffic monitoring is very important in addressing traffic problems.

With the fast increasing of transportation vehicles around the world, especially in urban areas, existing traffic management systems become inefficient need to be change. In this context, the technology advancement of wireless network communications and remote sensing, intelligent transport systems (ITS) have recently emerged as a key enabling technology to improve road safety, traffic efficienty and driving experience [1][2][3][4]. However, there are many open research challenges and issues that need to be tackled in order to foster the emergence of safe ITS applications. In this paper, we propose a scheme to automatically change and update the duration time of the traffic light which is being install “rigid” at intersections. Traffic data is collected by sensors embeded at the intersections, then classified to red or green cluster by applying Support Vector Machine. The duration of traffic light then is estimated by applying Webster calculation method. In the ITS including, the priority to update the traffic light state at each intersection is also considered by using Heuristic Intersection Selection Method and Intersection Variance Indicator. With our proposed scheme, the red/green signal light at the intersections will be updated automatically based on the real traffic data, and as a reference result for the management staff operating the transport system when necessary.

The remaider of this paper is structured as follows. Section 2 present our propose scheme. Section 3 shows simulation results and evaluation. Finally, Section 4 concludes the paper and provides future research directions.

2 A Scheme for Automatic Adjustment and Update Traffic Signal Duration (S-W-H)

In this section will describe the propose scheme in detail. The process of automatic adjustment and update the duration time of traffic light includes three main stage: Stage 1: Classify the traffic sensed data by applying Support Vector Machine (SVM) method. Stage 2: Calculate the duration time of red/green signal light using Webster's formular. Stage 3: Determine Intersection Variance Indicator (IVI) priority using Heuristic Intersection Selection Method (HISM) method, then selecting which intersection will be automatically updated the time.
Figure 1 shows the overall process of our proposed scheme

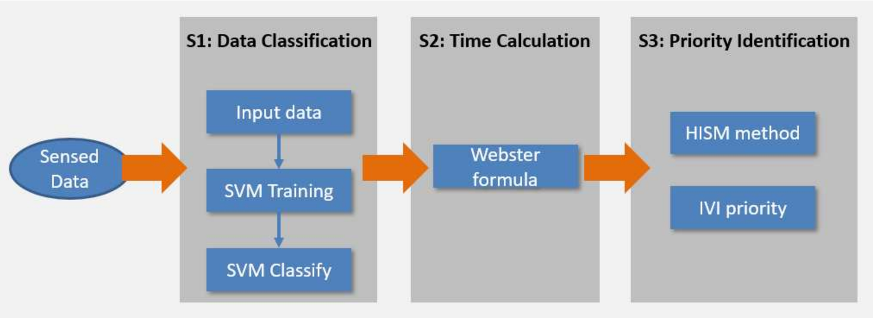


Fig. 1. The overal process

In this scheme, the input training data collected at the operating center will be trained by SVM training machine in order to classify into two clusters of red or green signal. The result will be the assessment set for the classification process of other data. In system real time, sensed data from the intersections will be classify by the SVM Classify learning engine in order to determine which cluster the current sensed data belongs to. In the next step, the current sensed data will be continuously processed in order to calculate the duration time for the traffic light. Thus, at the end of this stage, two kind of results are get: the state of traffic light (red or green) and the duration time of the traffic light. In the case of the intersections are interconnected, our proposed scheme provide a HISM mechanism to assess the appropriateness of each intersection. With HISM, the intersections will be calculate with an addition IVI value to assess the priority of the intersections after performing the time calculation stage.

2.1. Input data

Traffic data will be collected from two sources: sensor nodes embedded at traffic nodes, random data. In this paper, we consider the following traffic parameters: (1) number of vehicles, (2) type of vehicle, (3) speed of vehicle.

(1) Number of vehicles: This parameter plays a decisive role in calculatong the duration time of red/green signal of the traffic light. According to the regulations on traffic lanes in Vietnam for urban areas, each lane has a width of 3-8 meters. So, the number of vehicles at the intersections is relatively crowed.

(2) Type of vehicle: Most vehicles are divided into two types: non-priority transport vehicles (including cars, motobikes,…), and priority transport vehicles (including fire trucks, ambulances, military vehicles,…) which are allowed to continously move in case of the traffic light is in the red state.

(3) Speed of vehicle: In Vietname, the maximum speed in the densely populated area is from 50 km/h - 60km/h.

When a data stream is recorded, the system's database will save a set of three parameter (number of vehicles, type of vehicle, speed). On the other hand, intersections often have a number of lanes greater than or equal to 2. At that time, data stored at the center will be a set of parameters ([number of lane 1 cars, number of lane 2 vehicles , ..], [type of vehicle that prioritizes lane 1, priority vehicle of lane 2 ..], [speed of lane 1 car, speed of lane 2, ...]).

2.2. Data Classification

Nowaday, data classification are widely used in many different industries and research fields, such as: k-nearest neighbor (k-NN), neural networks, support vector machine (SVM),… Our goal is to determine the state of traffic light (red or green), SVM method is completely consistent to the problem. SVM is a binary classification method with two steps of training and testing. With a set of training templates in two given clusters, the SVM training algorithm builds an SVM model to classify other patterms into those two clusters. It classify the input data into two different clusters. The training process is based on the specific characteristics of the input data and select the appropriate boundary between the two clusters.

The classification of SVM consist of two form: 2-cluster distributed and multi-cluster distributed. It can be observed that 2-cluster distributed can be applied in the case of our proposed scheme. Then, the input parameter such as number of vehicles, type of vehicle, speed from the original data set will be used to perform as following: 1-Compare the type of vehicles at each lane: If a lane has any priority transport vehicle, it will determine the green classification. 2- In the absence of priority transport vehicles in both horizontal and vertical lanes, then perform a comparison on the number of vehicles per lane. The green classification is set for the lane with more number of vehicles. 3- If the lanes have the same value of two above parameters, the speed of vehicles will be used to determine the green classification.

2.3. Time Calcution

The method of calculation the optimal time-lapse cycle was built by Webster [5]. This Webster method is currently used by many country in calculating the duration time of traffic light. This method can be summarized in the three basic steps:

Step 1: Determine the average standby time of each transportation at the intersection by the following equation (1)

$$d = \frac{c(1-\lambda)^2}{2(1-\lambda x)} + \frac{x^2}{2q(1-x)} - 0.65\left(\frac{c}{q}\right)^{\frac{1}{3}} x^{2+5\lambda} \;.... \quad (1)$$

With:

d – Average waiting time of a vehicle (s);

c – Light cycle time (s);

λ – Ratio between effective green time g and light cycle time (λ=g/c);

q – Number of vehicles in 1 second, also called vehicles intensity (q=number of vehicles/s);

x – Saturation (x=q/ λs);

Step 2: Determine the total stanby time

$$D = \sum \textit{(d × number of vehicles)} \quad (2)$$

Step 3: Calulate the total standby time for the light cycle time, by solving the derivation equation (3):

$$\frac{dD}{dc} = 0 \implies C_o = F.\frac{2.L}{1-Y} \quad (3)$$

With:

C_o – Optimized light cycle time;

L – Total loss time in a light cycle time;

$Y = \sum y_i = \sum(q_i/s_i)$;

F – Coefficients determined by experiment;

q_i – Number of vehicles on the calculation lane of phase i;

s_i – Saturated level of number of vehicles on the calculation lane of phase i;

Webster used the theoretical equation (3) as well as empirical simulation to find out the optimal light cycle time by (4)

$$C_o = \frac{1.5\,L + 5}{1-Y} \quad (4)$$

However, applying equation (4) will not be optimal because of the transportation vehicles in Vietnam are not homogeneous. In order to solve that problem, we consider to apply another optimal equation (5)[6] for the heterogeneous transportation vehicles:

$$C_o = \frac{1.5L + 5}{1 - \sum \frac{1}{f}\left(\frac{q^{mc}}{s^{mc}} + \frac{q^{car}}{s^{car}}\right)} \quad (5)$$

With:
C_o – Optimized light cycle time;
L – Total loss time in a light cycle time;
f = 1
q^{car} and q^{mc} – Number of cars and motobikes
s^{car} and s^{mc} – Saturated level of number of cars and motobikes

Then, the duration time for green light is calculated as follows, equation (6)[6]:

$$G_i = \frac{y_i}{\sum_{i=1}^{n} y_i}\, x\, (C - L) \quad (6)$$

In summary, from studies and distributions about Vietname traffic, the red/green signal cycle and green signal duration time are calculated with constant of parameters, such as: L = 5; f = 1; s^{car} = 35; s^{mc} = 30; q^{car} = number of cars; q^{mc} = number of motobikes. Then, equation (5) and (6) are rewrite as follows:

$$C_o = \frac{1.5 * 5 + 5}{1 - \sum\left(\frac{q^{mc}}{35} + \frac{q^{car}}{30}\right)} \quad (7)$$

$$G_i = \frac{y_i}{\sum_{i=1}^{n} y_i}\, x\, (C_o - L) \quad (8)$$

2.4. Priority Identification

At every traffic node, IVI parameter is calculated using equation (9). Descending sort IVI values at intersections. The intersection which has the largest IVI value will update the duration time of red/green signal light. The remaining unmodified intersections will change the duration time by reinstalling previously installed values when the timw is over.

$$\text{IVI} = \frac{\sum_{i=1}^{m} \Lambda_i \beta_i}{m}\ , \beta_i = \frac{1}{2s}\left[\frac{G}{R+G} - \frac{f}{s}\right]^{-1} \quad (9)$$

With:
m – Number of lanes at an intersection
Λ_i – Average waiting coefficient at an intersection
G – Green signal duration time
R – Red signal duration time
s=1
f=1

3 Simulation Results and Evaluations

All experimental simulations are run on computer systems with basic hardware components including: Intel® Core I7 -3520 2.9 Ghz 4 Core CPU (4 threads), 8GB Ram. All algorithms are coded to run on the version Matlab 2017b software including the main functions as follows: Conduct random data creation, perform data classification, conduct temporal calculations red/green signal light, calculate HISM priority index at each intersection,... These processing functions are written separately into files according to the structure of the Matlab 2017b application.

Figure 2 shows the system interface including four intersections (traffic nodes). Each intersection including the following information:

- Green bar: represents the green light
- Red bar: represents the red light
- A sequence of three numbers: Total time of red/green light, the number of vehicles passed the intersection, the time of red/green light has passed
- Red square: repesents a car (equals 4 motobikes)
- Blue square: represents a motobikes

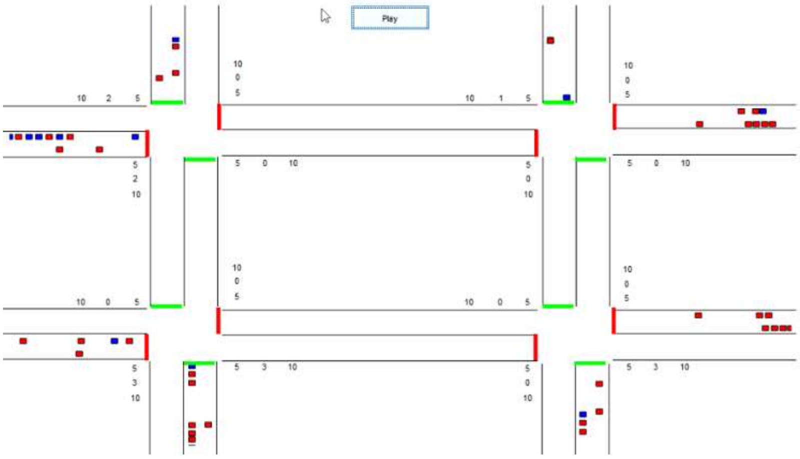


Fig. 2. System interface

Initially, the sequence of three numbers at the intersections are set by 0. The system randomizes the data, calculates the duration time of red/green light of the intersections. Then, the system operates scanning the environment in order to detemine the numbers of vehicles and calculate tien signal time. The state of signal lights as well as their duration time are realtime updated.

During simulation time, our propose can effective solve almost normal cases of a traffic system as well as some special cases. Figure 3 shows a case of "An intersection with signal time is over" (Intersection #1) and of "An intersection with the largest IVI index, but signal time is not over" (Intersection #3)

At intersection #1 has some special characteristics:

- Duration time of signal lights at both lanes are over

- When the time is out, the state are reset and value of red/green signal lights are set as initial values
- The system represents with the yellow block in order to notice to the administrator whether this case happens.

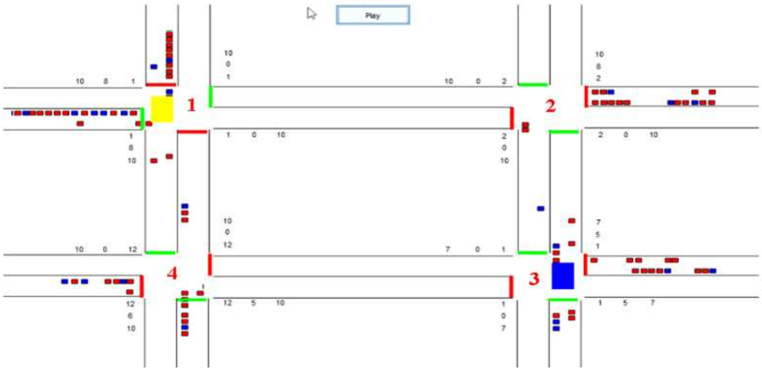


Fig. 3. A case of "An intersection with signal time is over" (Intersection #1) and of "An intersection with the largest IVI index, but signal time is not over" (Intersection #3)

At intersection #3 has some special characteristics:

- IVI index is largest
- Duration time of signal lights are not over, vehicles are moving. Thus, the intersection itself update the signal light state and value with the previous value.
- The system represents with the blue block in order to notice to the administrator whether this case happens.

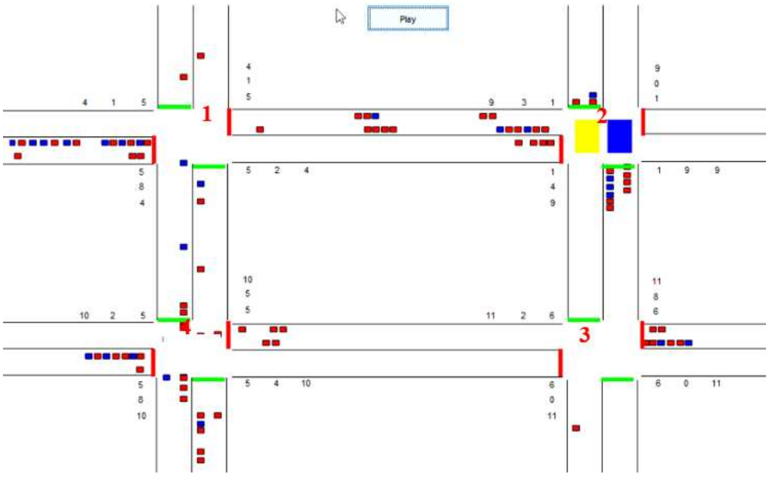


Fig. 4. An intersection with the largest IVI index and duration time of signal lighs are over

In case of Figure 4, “An intersection with the largest IVI index and duration time of signal lighs are over”. We can observe that the yellow block and blue block are located at the intersection #2, the system has some special characteristics:

- IVI index is largest
- Duration time of signal lights at both lanes are over
- The system operates calculating the state and value of signal lights using current information at this intersection

The system represents with the yellow and blue block at the same intersection in order to notice to the administrator whether this case happens

4 Conclusion and future works

In this paper, we have proposed a effective scheme to automatically change and update the duration time of the traffic light which is being install “rigid” at intersections. We have used Matlab to do simulate a traffic system. The system has run effective and solve almost cases may happen in a traffic system.

In future works, we are going to apply other diffirent classification methods (such as k-nn, neutral network,…) combinationing with Webster and HISM algorithms to get better comparision results.

References

1. Hamida E.B., Noura H., Znaidi W, “Security of Cooperative Intelligent Transport Systems: Standards, Threats Analysis and Cryptographic Countermeasures”, *Electronics* ,vol. *4*, 380-423, (2015)
2. Mohamed Amine KAFI et al., “A Study of Wireless Sensor Networks for Urban Traffic Monitoring: Applications and Architectures,” in The 4th International Conference on Ambient Systems, Networks and Technologies(ATN2013), Halifax Nova Scotia, Canada, 2013, pp 617-626
3. Ondrej Karpis, “Wireless Sensor Networks in Intelligent Transportation Systems,” International Journal of Modern Engineering Research (IJMER). Vol 3, Issue.2, University of Zilina, Slovakia, March-April. 2013, pp 611-617
4. Malik Tubaishat et al, “Wireless sensor networks in intelligent transportation systems,” in Wireless Communications and Mobile Computing, 2009. Wiley InterScience. Vol.3 Columbia, U.S.A, 2009, pp.611-617
5. Webster, F.V (1957), Traffic Signal Settings, Road Research Technical Paper No.39
6. Do Quoc Cuong, “Traffic Signals in Motorcycle Dependent Cities.”, Technische Universität, Darmstadt, [Ph.D. Thesis], (2009)

Interactive 3D Visualization of Smart Manufacturing Data

Bakhit Sadirbaev[1], Rockwon Kim[2], Aziz Nasridinov[1], Kwan-Hee Yoo[1*]

[1]Dept. of Computer Science, Chungbuk National University, South Korea
[2]Intelligent Cognitive Technology Research Department ETRI, Daejeon, South Korea
sadirbaev@chungbuk.ac.kr, rwkim@etri.re.kr, aziz, khyoo{@chungbuk.ac.kr}
*Corresponding Author

Abstract. Target paper aims to provide the Data Driven Visualization (DDV) library, especially for smart manufacturing systems. A big amount of data can be illustrated conveniently in single screen by the combination of 3D and 2D charts. As well as, mouse events give interactive opportunities to stockholders. Precisely, zooming, dragging, hovering and the most significant feature is after clicking an item, drawing another chart by pre-designed scenario. Developer-friendly library was developed with JavaScript and it can be easily implemented to Web applications.

Keywords: 3D, 2D, visualization, smart manufacturing systems, DDV.

1 Introduction

Countless data is generated in Manufacturing companies. However, these data are useless if they are not presented to the right people, at the adequate time [1]. Prior to last three decades, 2D visualization was de facto standard to visualize data with such graphs as a line graph, bar graph and others [2]. As well as, 3D visualization gives a myriad of opportunities in the field of data science and big data analytics. We aim to develop a new Data Driven Visualization (DDV) library to visualize 2D and 3D graphs on a single screen simultaneously. The paper is organized in the following way: after the introductory paragraph we cover related studies of the fellow researchers and their gained experience. Section 3 will be about the general description of the proposed DDV method and its efficiency. Next section provides information on the results of the conducted research and it is followed by suggested points for the future experiments on this area as the final part.

2 Related Works

Nowadays, big data analytics has been using in manufacturing industry to support its research and operational activities [3]. So far, a few researches have been done in this area, Jo et al. [4] extend the basic Gantt chart for the exploration of large schedules. Worner and Ertl [5] propose a novel visual analytic system for simulated manufacturing processes. These studies visualize the data related to the planning and simulation stages in manufacturing. Xu et al. [6] described the design of a visual analytic system for manufacturing process data collected during the operation of the assembly lines in modern factories. They proposed design and implementation of a comprehensive visual analytics system, ViDX. It supports both real-time tracking of assembly line performance and historical data exploration to identify inefficiencies, locate anomalies, and form hypotheses about their causes and effects. We also tried to correspond applicably to the demands of the stakeholders in Manufacturing Systems occasionally. Furthermore, three-dimensional charts, considered to be a major idea of our recently established library, are excluded in the previously mentioned tools or libraries.

3 Proposed Methods

In visualization area, data is a top priority. There are number of data serialization formats, however nowadays JSON (JavaScript Object Notation) is widely used [7]. Therefore, JSON has been used as input data format. DDV library is based on Three.js. It has all required components like scene, camera, lights, box, text, line etc., to illustrate data as 3D graph. Visualization tools should be interactive, and user engagement is very important [8]. In order to make it more interactive, OrbitControl.js, which is extension library to three.js, is used. Orbit controls allow the camera to orbit around a target. Precisely, stockholders can drag, rotate and zoom the graph. As well as, mouse events, like hovering and clicking, also give more opportunities to users. There is also open source extension library for mouse events. threex.domevents provide dom events inside 3D scene (Fig. 1).

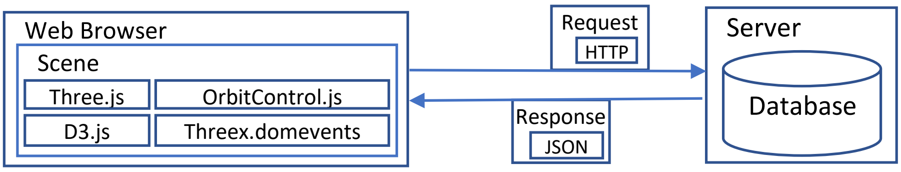


Fig. 1. Flow of the 3D Visualization process.

Through hovering an item, a brief information is shown on the top corner of scene. Additionally, if you click an item, another 3D graph is drawn with related information of clicked item. For this, special scenario is created. As well as, if you click right button of mouse, context menu is shown. There are two items, they are for drawing 2D graph daily or weekly. For 2D graph, D3.js is used. The actions are shown completely in Table 1.

Table 1. Instruction of available actions.

Action	Instruction
Zooming out	Scroll up mouse wheel on the scene
Zooming in	Scroll down mouse wheel on the scene
Dragging	Click right button and move the mouse on the scene
Rotating	Click left button and move the mouse on the scene
Hovering	Move mouse pointer on items
Clicking	Move mouse pointer on item and click left button of the mouse
Context menu	Move mouse pointer on item and click right button of the mouse

4 Experiment Results

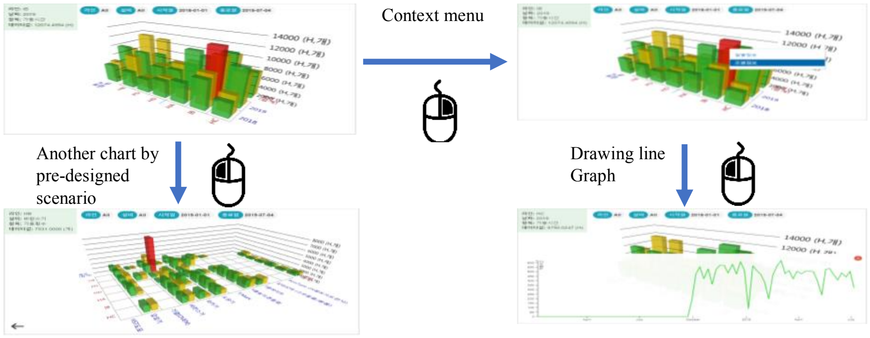


Fig. 2. The result of 3D visualization library with default settings.

DDV library is especially for web development, for this reason, html, CSS and JavaScript programming language were used to develop this library. First, all required libraries and ddv.js should be imported. They are three.js, OrbitControls.js, threex.domevents.js and D3.js. Input data format should be JSON, and item colors and titles need to be given. Input data was uploaded to Github as a sample, to get more information, follow this link: https://github.com/qarakenbacho/DDWV/blob/master/input_data.json

During initializing DDV, id of container, which visualization is drawn in, need to be shown. After calling `barChart()` function, the input data visualized with 3D graph with default settings (Fig 2).

5 Conclusion

In this paper, we explored visualization library, which consists of 3D, 2D charts and other interactive actions. Precisely, this library gives opportunity to visualize data with combination of 3D bar chart and 2D line chart. This kind of libraries allow to users to make better decision in a short time, if it uses in a right way. Three-dimensional bar chart is good to visualize and simple to understand, however, our plans for future are implementing other types, like 3D pie charts, 3D scatter plots and 3D bubble plots. As well as, enriching with animations is also make this library more interactive. On the other hand, the most vital sides of visualizing are speed performance and meaningfulness [9]. In the future, we will do research about the fastest visualization tools and learn their experience. Purpose is developing rich and powerful 3D data visualization library, which is answerable to developed world requirements.

Acknowledgments. The work is supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under i-Ceramic manufacturing innovation platform technology development business. No.2004367, 'Development of cloud big data platform for the innovative manufacturing in ceramic industry', and by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW (2019-0-01183) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation).

References

1. The 8 Best Data Visualization Tools, https://www.bernardmarr.com
2. Ulrich, L., Joachim, K., Uwe, W.: 3D Visualization and Animation – An Introduction. In: Dieter, F. (ed.) Photogrammetric Week '03, pp. 217--226. Wichmann Verlag, Heidelberg (2003)
3. Auschitzky, E., Hammer, M., Rajagopaul, A.: How big data can improve manufacturing. McKinsey & Company (2014)
4. Jo, J., Huh, J., Park, J., Kim B., Seo, J.: Livegantt: Interactively visualizing a large manufacturing schedule. IEEE Transactions on Visualization and Computer Graphics, vol. 20, pp. 2329--2338, (2014)
5. Worner, M., Ertl, T.: Visual Analysis of Advanced Manufacturing Simulations. In: Miksch, S., Santucci, G. (eds) EuroVA 2011: International Workshop on Visual Analytics, The Eurographics Association (2011)
6. Xu, P., Mei, H., Ren, L., Chen, W.: ViDX: Visual Diagnostics of Assembly Line Performance in Smart Factories. IEEE Transactions on Visualization and Computer Graphics, vol. 23, pp. 291--300 (2017)
7. Douglas, C.: The JavaScript Object Notation (JSON) Data Interchange Format. In: Tim, B. (ed) (2017)
8. Simon, P.: The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions. Harvard Business Review, pp. 1-8. (2014)
9. Marius, G., Nur, B.M., Julian, H., Daniel, P.: Digital Twins in the Smart Factory. Journal of Engineering, Management and Operations Vol.I, pp. 41—54. (2018)

A platform for exposure assessment survey data

Hye Jin Lim, Na Ri Park, Yeong Hyeon Gu, Seong Joon Yoo

Departments of Computer Engineering, Sejong University,
209, Neungdong-ro, Gwangjin-Gu, Seoul, Korea
yoon1004@gmail.com nari.park@sejong.ac.kr
yhgu@sejong.ac.kr sjyoo@sejong.ac.kr

Abstract. Among the many technical and procedural challenges in exposure assessment, product usage data collection would be undoubtedly one of the most crucial components of research for its quality. Many cross-sectional studies on evaluating exposure to household products conducted surveys on the aspects of product usage and environmental factors. Due to the nature of such research, the sample size should be sufficiently large enough to consider the results would not be seen by chance alone. Also, data on exposures and outcomes must be captured for consistently and continuously to achieve the objective of the research, monitoring individual's integrated exposure by different exposure scenarios and routes of exposure. However, it all has implications for cost. In this paper we introduce a comprehensive platform consisted of CAERIG (Chemical Aggregate Exposure Risk Information Guide) for data collection and analytics on exposure to household products, CAERIG COMMUNITY for candidly exchanging post-purchase experiences among users, providing an interactive format of the information on products and potential exposure risk for users, and finally, making available ready-to-review exposure data for experts, manufacturers, and policy-makers.

Keywords: Household Products, Chemical Exposure Assessment, Survey Data, Visualization, Deep Learning, Python, CAERIG

1 Introduction

Exposure risk from using everyday products is one of the most serious social issues in South Korea in the past decades, represented by two major toxic scandals, the deadly humidifier disinfectant of Oxy and toxic sanitary pads. The disinfectants were first introduced to the market in 1994, with more than 20 different types being sold until 2011. Over the 18 years, eight million people are presumed to have used the product and about 130 people have been officially confirmed to have died from using the disinfectants. Although the suppliers and the manufacturers bear the primary responsibility for the tragedy, the absence of an integrated legal and administrative government body is also to blame. To that end, the Act on the Registration and Evaluation of Chemical Substances of Korea, known as K-REACH, was promulgated in 2018 and already came into force as of 2019. Under the Act, the designation of hazardous chemical substances must be regulated through registration and evaluation, for cleaning agents, synthetic detergents, bleaching agents, fabric softener, coatings, adhesives, fragrances, and deodorants to start.

Nevertheless, this Act alone would not be able to prevent and control exposure risks. While the most common chemical substances used in household products will be subject to registration, some are manufactured using toxic chemicals in a process known for its hazardous byproducts. Exposure risks may depend on individual susceptibility and real-life use scenario. Negligence of manufacturers or regulatory body should not be overlooked, and we cannot leave it to chance alone. Public attention and engagement are vital sources as seen in toxic sanitary pad scandal. Online communities had not only recognized the issues even before the national outbreak but created initiatives at grassroots level within the communities. However, they could neither act as a public alert system at the time, nor spread countrywide in timely manner because such online communities are mostly private and open to registered members only. Only if the issue was openly discussed with a greater level of support of regulatory system on time, the damage would have been much less, socially and economically.

The role of ICT is emphasized more than ever as a means of drawing social attention and engagement with existing and emerging social issues. Hence that the Ministry of Science and ICT of South Korea has planned and directed "living lab," a new R&D model that suggests a collaboration of public-private-people partnership, allowing all involved stakeholders to work together in discovering ways in which social issues might be solved. As part of its R&D plan, we developed a mobile-based, comprehensive platform, CAERIG, or Chemical Aggregate Exposure Risk Information Guide, to be a risk communication channel that reaches the entire society from raising public awareness and assessing the risk from everyday products to assisting policy makers with designing and refining new policies and regulations in real-life scenario before implementation.

Section 2 introduces the objectives and potential benefits of the project and section 3 explains the architecture of CAERIG and CAERIG PRO. The details about the platform such as user interface, visualization, exclusive features are described in section 4. Section 5 discusses the effectiveness of the CAERIG as applied to real-life cases and opportunities for improvements.

2 Background

Exposure assessment has shifted from environmental factors toward everyday products with personal exposure monitoring. Such trend has raised consumer awareness and ethical and/or scientific challenges. It requires the public be involved in the research, experts interpret and visualize the findings in layman's term for consumers, and eventually, new social atmosphere be created toward chemical household products. The CAERIG described in this paper is designed specifically for: collection of exposure-related data from the public through mobile-based web application, easing the burden of participants; informing the consumers of their product usage patterns and exposure risk along with substances information without jargon; sharing post-purchases experiences among consumers; providing an opportunity for policy makers, researchers, and manufacturers to explore public views and perspectives. Public engagement with research through CAERIG will generate mutual benefit between the public and researchers, and ultimately, enhance the quality or impact of research.

3 Exposure Assessment System Framework

Exposure assessment system consists of the followings: data platform, calculation engines, and reporting platform [Figure 1]. CAERIG data platform collects only the minimum amount of information necessary for exposure risk calculation, namely, age, weight, product in use, frequency, and quantity per use from the users. Based on the substances, constants, and usage scenario, individual's risk characterization ratio (RCR) is calculated and stored until further update by the user. The risk is interpreted and visualized in user-friendly knowledge base forms.

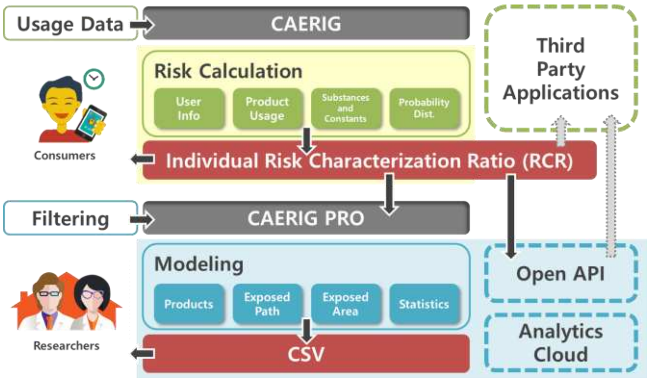


Figure 1. CAERIG Framework for consumers and researchers

Researchers and experts who need datasets or references for their research filter for products, exposed paths, or exposed body parts on CAERIG PRO data platform. Modeling module then statistically analyzes the datasets from CAERIG and return the result by criteria in a comma separated values file. Open API and a remote analytics cloud services will soon launch once de-identification procedures complying with the current regulatory standards applied. New services or improved existing services are expected in the market by third-party applications.

3.1. CAERIG

CAERIG project aims to build a database of ingredients and toxicity for chemical substances in domestic chemical products and develop big data analytics to provide information on products, substances and toxicity, personal exposure, risk assessment, and usage monitoring. Public's lack of understanding about chemicals, media coverage and product marketing messages leading to misperceptions, irrational fear, and an inability to correctly determine risk affect the misuse of products or needless anxiety over chemical substances. As part of CAERIG project, researchers have conducted scientific surveys of 10,000 samples to measure current attitudes and behaviors about chemical substances and product usage.

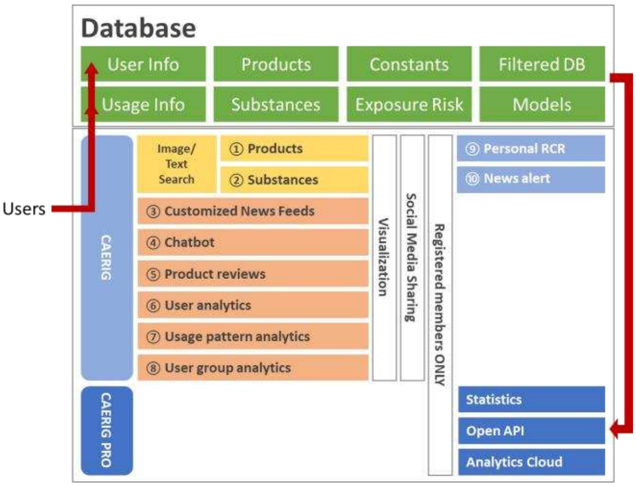


Figure 2. Features of CAERIG

One of the CAERIG project's goals is the collection, storage, sharing, and dissemination of biophysical and social datasets. We have generated baseline biophysical datasets and a mobile-based platform to collect usage data, interpret and visualize the exposure and risk, and share the findings and datasets through Open API and remote analytics clouds for researches and third-party services. For data collection and visualization, the target audiences include any consumer who uses household products and wants to monitor one's usage behaviors [Figure 2].

As in a case of survey data for the baseline dataset, we made a tab called "Diary" to collect everyday use patterns of household products, implemented by participating researchers. The primary objective of the survey was to document how a representative cross-section of the population is using, for example, how much and often, whether using the products as directed with or without protections and ventilation, and so on.

The survey was administered by the department of Bio-Convergence from Sejong University collaborated with Consumer Organization. It had been implemented online collected 10,000 qualified responses for 8 product groups, total of 189 products. The survey will be on-going as soon as "living lab" feedbacks and suggestions are applied. Also, the survey was designed to contain no personally identifiable information, so sharing of the dataset would not reveal the identification for respondents.

3.2. CAERIG PRO

Considering a broad audience with varying levels of technical and scientific expertise, we developed CAERIG PRO, a web-based statistical modelling tool, to provide pre-defined statistics on exposure and risk based on user's own criteria [Figure 3]. As soon as the range and policies on data dissemination are settled by participating researchers, an Open API and a remote analytics cloud would be open to public to provide an access to survey datasets. Unlike CAERIG, CAERIG PRO service may be limited to registered experts, policy-makers, and manufacturers who have interests in norms and standards in household product exposure and risk.

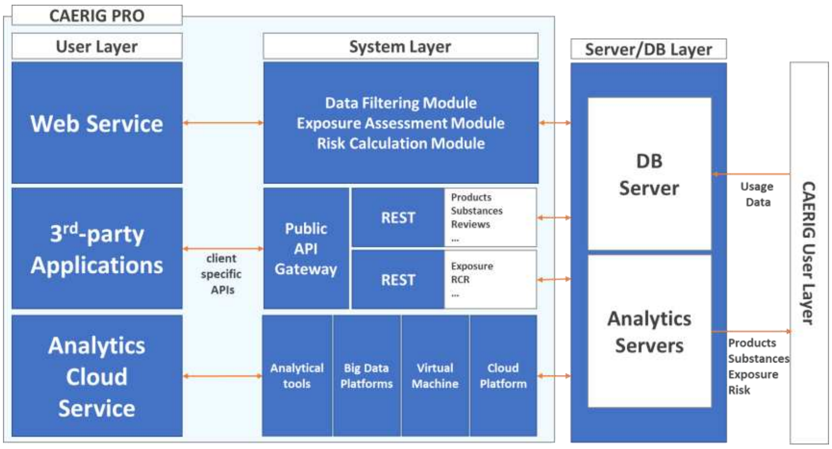


Figure 3. Open API and Cloud Architecture of CAERIG PRO

The analytics cloud is built on OpenStack™ with most frequently used analytical environments such as R, Python, and more. As a virtual machine and big data platform, it provides a commercial open source framework for real-time, advanced analysis including high-speed query execution tools, machine learning libraries, graph processing and streaming engines. It is also tool for various big data analysis and utilization technology development and eliminates unnecessary instance regeneration and resets analysis and adds ease of use. Registered users will be given an access to data for analyzing usage patterns and review data collected from CAERIG to de-identified data and analyzing big data related to the usage of chemical products.

4 Mobile Web Application Implementation

The major goal of the CAERIG was to present simple, interactive, and easily comprehensible visualization of quantitative and qualitative survey result and one's own exposure assessment. The architecture of the web application consists of user interface, web framework, analytics layer, and a data storage layer [Figure 4].

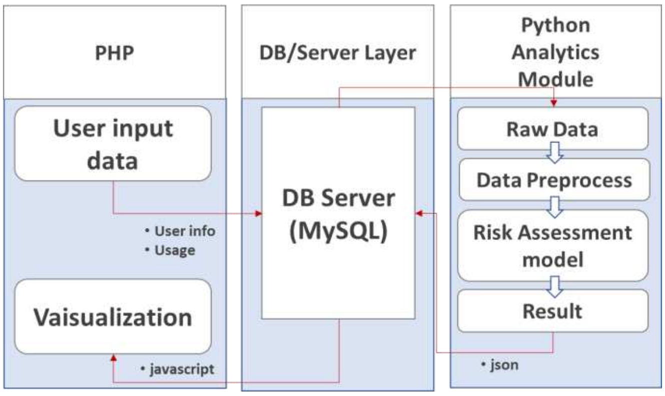


Figure 4. Architecture of CAERIG mobile web application

4.1. User Interface

The user interface was implemented using HTML5 and JavaScript, making the CAERIG cross-platform compatible for users on any device with a mobile web browser. For the user interface design and usability, we used the Bootstrap framework, which provides effective and responsive user experience along with consistency for all types of browsers.

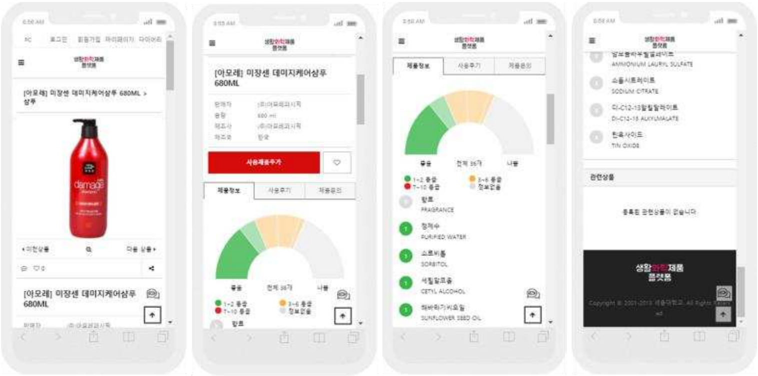
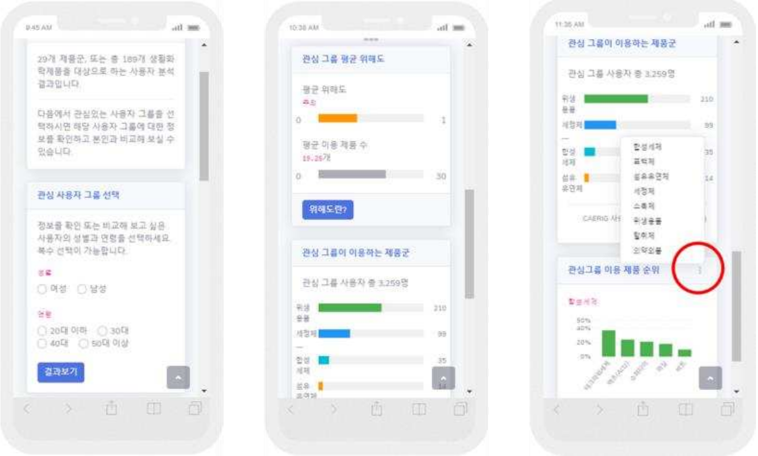


Figure 6. Product and ingredients information pages

The user interface consists of a landing page where users can search for products or substances by either text or image, personal page where users can alter their information, and diary page where users submit their daily use of products. It also has "About" page where RCR and RCR calculation are explained, product and substances information page, and community page where users can aggregate and disaggregate of survey responses and product review feeds based on demographic variables or similar user groups. They will be integrated into CAERIG platform as soon as living lab feedbacks and suggestions are applied. Figure 6 shows product and ingredients information and Figure 7 show the demo pages of interactive user group usage patterns and RCR information page that are currently under living lab's review.



29개 제품군, 또는 총 189개 생활화학제품을 대상으로 하는 사용자 분석 결과입니다.

다음에서 관심있는 사용자 그룹을 선택하시면 해당 사용자 그룹에 대한 정보를 확인하고 본인과 비교해 보실 수 있습니다.

관심 사용자 그룹 선택

정보를 확인 또는 비교해 보고 싶은 사용자의 성별과 연령을 선택하세요. 복수 선택이 가능합니다.

성별

○ 여성 ○ 남성

연령

○ 20대 이하 ○ 30대
○ 40대 ○ 50대 이상

결과보기

관심 그룹 평균 위해도

평균 위해도
주의

평균 이용 제품 수
19.26개

위해도란?

관심 그룹이 이용하는 제품군

관심 그룹 사용자 중 3,259명

위생용품	210
세정제	99
합성세제	35
섬유유연제	14

Figure 7. Disaggregate response of survey pages

4.2. Web Framework

The CAERIG uses the Apache and PHP web framework as a mediator between the front-end layer and the underlying web server. To have all the configuration settings for the PHP engine, the Apache HTTP Server, and the MySQL database server specified automatically, use an AMP package [Figure 8]. A separate Django server is running for statistics data generation for JavaScript data visualization, csv export at CAERIG PRO, and Open API.

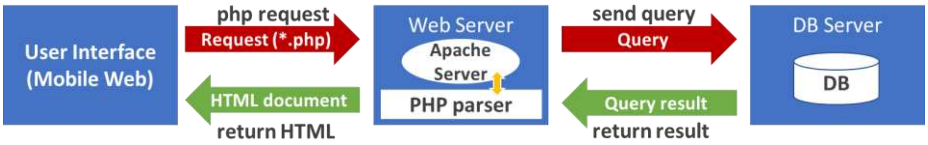


Figure 8. PHP architecture overview

4.3. Data Storage

PHP object relational mapper (ORM) that sits on top of a powerful database abstraction layer (DBAL). One of its key features is the option to write database queries in a proprietary object-oriented SQL dialect called Doctrine Query Language (DQL). This provides developers with a powerful alternative to SQL that maintains flexibility without requiring unnecessary code duplication.

The data file is a table that consists of a single column for each survey question, a variable, and a row for each respondent, a consumer. Search logs will be stored for further issue detection analytics once open to public. For quantitative responses regarding usage patterns, they are used in exposure assessment and risk calculation, and for the qualitative responses such as product review feeds and chatbot log data are analyzed through text mining and natural language processes with python modules.

4.4. Big Data Analytics Features

The platform has several big data analytics functionalities for the sake of users in forms of user interface and information page. As shown in Figure 9 and 10, we implemented a similarity-based image search module for users to easily search products with images without texting product names. We have trained with 3,353 images for 21 different shampoo brands and 1,440 images for 17 different rinse brands. Images were trained with a convolutional neural network model called VGG16-FC1, using default ImageNet weights, after removing the background. Next, the retrieved images are analyzed using object detection and the ranking of results are adjusted according to distance from a cluster representative to a query. VGG16 was trained for days and was using NVIDIA Titan GPU's.

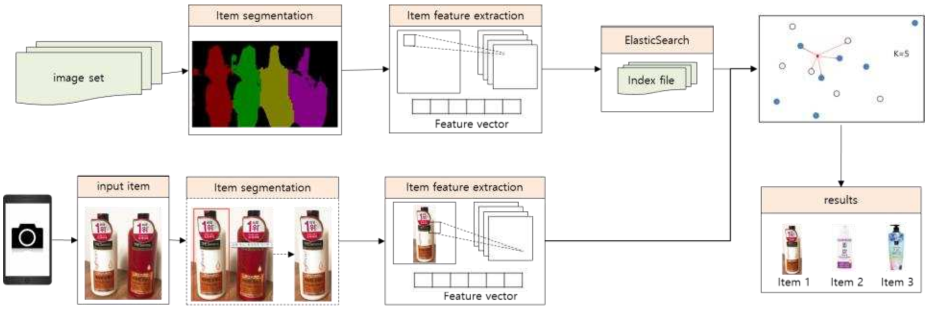


Figure 9. Deep Learning based product search system

Natural language process has been applied to the reviews scraped from online shopping sites since it's not open to public yet, there is no reviews made. The purpose of product review is to detect issues in advance and prevent any incidents as early response to issues can be made on time. As shown in Figure 11, reviews are shown in ratings and number of users who reviewed for each product group, daily sentiments, and keywords both in reviews and search.

Since the images are of prepared in various situations in terms of contrast, composition, background, the accuracy and precision are to be measured after the platform is open to public receiving users' actual photo images. It applies to product review analysis as well for that the analytical model is trained by and developed from using online shopping sites' reviews, which consists of other factors than experiences regarding safety, such as delivery, price, scent, cleansing, customer service of the website, and so on.

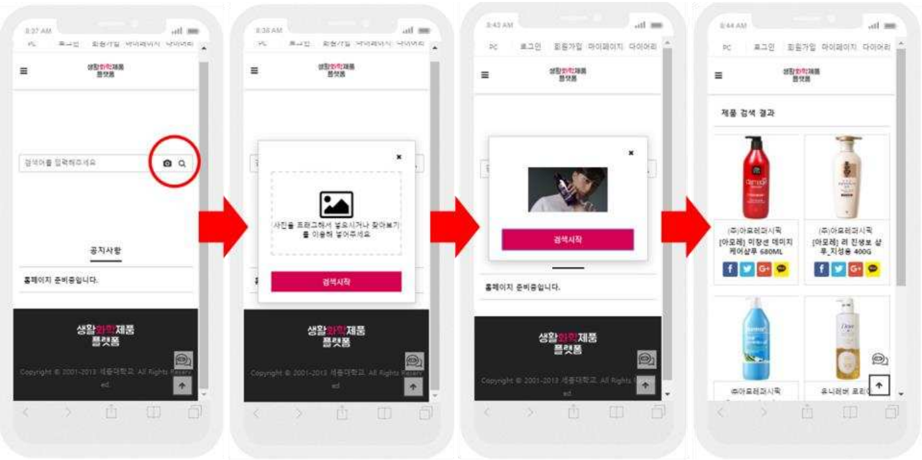


Figure 10. Image Search Example

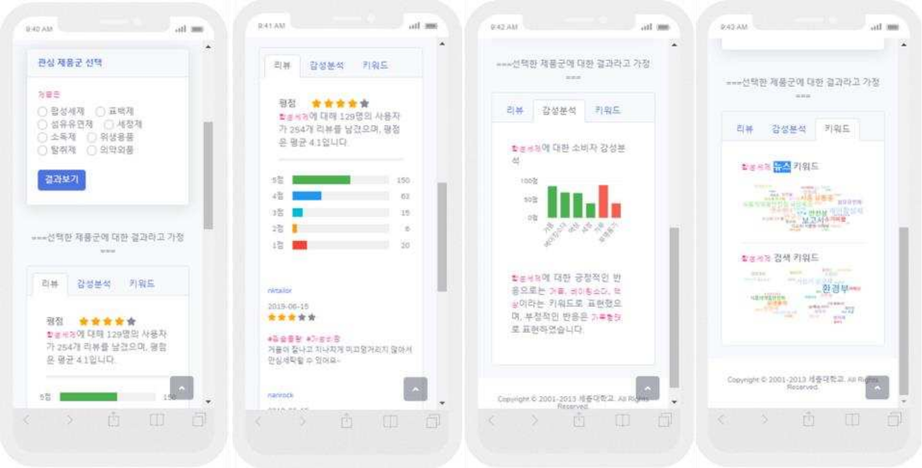


Figure 11. Examples of product review data analytics page

5 Conclusion

The CAERIG provides a mechanism for making quantitative and qualitative survey data accessible to consumers and a broad range of technical and scientific experts. The mobile-based interface that requires no other software for users to submit information and visualize the exposure assessment. In addition to personal monitoring, it also displays responses to various survey questions by percentage or mean results by similar user groups if disaggregated by demographic variables. CAERIG PRO provides easy access to datasets and analytical environments for policy makers, researchers, and manufacturers.

Although it is not without its limitations, it contributes a potentially valuable tool that facilitates the mobile-based interactions between public and regulatory system of government and manufacturers. The CAERIG allows stakeholders to explore household products usage data and their own exposure assessment by monitoring everyday use. We expect many third-party applications use our Open API service to create consumer-centered services, as a social surveillance on hazardous everyday products. Although the CAERIG focuses on chemical exposure assessment, the features and functionalities could be applied to other social domains in promoting public to participate in resolving social issues.

Acknowledgements

This research was supported by Research Program to Solve Social Issues of the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT. (NRF-2018M3C8A6081624).

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00136, Development of AI-Convergence Technologies for Smart City Industry Productivity Innovation)

References

1. Robinson, L. (2018) “A Practical Guide to Toxicology and Human Health Risk Assessment”. WILEY.
2. Gautsson, T.; Larsson, J.; Staron, M. (2015) "Collecting Product Usage Data Using a Transparent Logging Component". International Conference on Advances and Trends in Software Engineering.
3. Frommea, H.; Albrechtb, M.; Angerer, J. (2007) " Integrated Exposure Assessment Survey (INES): Exposure to persistent and bioaccumulative chemicals in Bavaria, Germany". International Journal of Hygiene and Environmental Health Volume 210.
4. Heinälä, M.; Gundert-Remy U.; Wood, M. H. (2013) “Survey on methodologies in the risk assessment of chemical exposures in emergency response situations in Europe”. journal of Hazardous Materials.
5. Arnold, S. M.; Greggs B. (2017) “A Quantitative Screening-Level Approach to Incorporate Chemical Exposure and Risk into Alternative Assessment Evaluations”. Integrated Environmental Assessment and Management.
6. @fchollet, The Keras Blog (2016) “How convolutional neural networks see the world [Web blog post]”. Retrieved July 14, 2019 from https://blog.keras.io/how-convolutional-neural-networks-see-the-world.html

Development of Logistics in Uzbekistan: Current State and Application of Big Data Technologies

Azizbek Marakhimov,[1], Abdulaziz Mavlyanov[1],
[1] INHA University in Tashkent, School of Logistics, 9 Ziyolilar Str, Tashkent, Uzbekistan
a.marakhimov@inha.uz

Abstract. This paper reviews the current state of logistics development in Uzbekistan. As of one of the landlocked “Stan” countries, Uzbekistan has been concentrating its effort in fostering the logistics sector. One of the promising projects in this vein was logistics center built in Navoi city. Recent observations indicate government’s dissatisfaction with the Navoi logistic center’s performance. In this paper, we review key flaws that lead to the low performance of the center. Particularly, we pay attention to the improper digital strategy implemented by the management of the Navoi logistics center that omitted Big data initiatives.

Keywords: Review, Big data, Logistics, Central Asia, Sustainable development

1 Introduction

Logistics is an inseparable part of a sustainable development of a country. It is a constantly evolving field of science that is tightly integrated with many other disciplines, such as marketing, information technology, and economy. Hence, there is no single definition for logistics. However, Kovács (2016) has managed to grasp its concept by stating that “logistics is the planning, organizing and coordinating of the flow of materials, information, energy, money and values… logistics is also an interdisciplinary discipline that synthesizes and utilizes the state-of-art knowledge and methods of several disciplines.” (Kovács, 2016).

In Central Asia, particularly in Uzbekistan, the development of logistics is at a high priority. By liberalizing their markets, the “Stan” countries have taken the burden of creating efficient supply chain networks. Many impediments in the form of legal, cultural, technological, and institutional issues stand in their way. A major one is the practical undervaluation and infrequent use of modern analytical tools, like Big Data. Lack of these vital decision making instruments in modern logistics operations has negative implication on the performance of organizations. Such an incident took place in Uzbekistan and involved the failure of the new Navoi Logistics Center (LC). We are particularly interested in the reasons why the LC failed despite continuous efforts from both the South Korean and Uzbek governments and how it could have been prevented with the help of Big Data. The results of this analysis could help the LC identify its core problems and hopefully recover in the future, and also serve as a case study for other LCs that may be facing a similar situation.

2 Distribution and Logistics Development in Uzbekistan

Development of logistics correlates with the changes in the retail sales (Fernie and Sparks, 2014). After the global financial crisis in 2008 the world slowly started to recover (Fig 1). Now the global retail industry is seeing positive changes with sales exceeding $24 trillion in 2018, which in turn is helping the logistics sector grow and prosper (Statista, 2019). In turn, Central Asian countries including Uzbekistan are concentrating their effort to develop logistics sector.

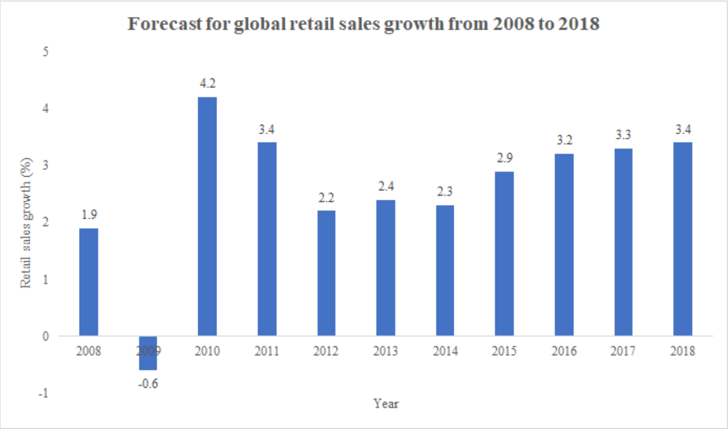


Fig. 1. Forecasted growth of global retail sales (Statista, 2019)

2.1 Establishment of Navoi Logistics Center

The Navoi Logistics Center (LC) has a long history. Starting as an airbase in the very beginning, the Navoi LC has become what it is due to relatively recent events. In 1962 the Navoi airport was built, and after its modernization in 2007 it slowly began to warp into the Navoi LC known today. The Uzbekistan's coop with Korean Air played a huge role in the development of the LC. From 2009 to 2013 Navoi LC was under construction. And in June 2013 the LC was officially launched. At that time, it was created to perform a number of operations, such as handling of multiple types of cargo, warehousing for cargoes with specific requirements, cargo transit services, and information services. The LC was destined to become efficient cargo processing location with high quality standards, best technology and equipment, and skilled personnel.

There are a number of reasons why Navoi was selected as the site for the LC. Location is perhaps the most important reason why. A number of big highways, railways, and airway corridors pass through the Navoi centre. So far there is no direct connection between the west and east. That is why so many airlines are interested in cooperating with Navoi. For instance, in the world map of air routes, the link between Moscow and New Delhi passes through the bottom of Uzbekistan where Bukhara and Navoi are located. Thus, certain flights from Europe and East Asia can utilize the Navoi

LC as an air transit point. The first President of Uzbekistan, Islam Karimov, stressed that Navoi LC and the Free Trade Zone (FTZ) are located in a very strategic location. The two objects, as well as the cargo center (TIR), are closely located to each other, enabling high cooperation. (Hudoynazarova, 2015) The Uzbek government hoped that the centre will serve as air transit point connecting Europe with East Asia and promote the export of the nation's industrial and agricultural production.

2.2 Navoi LC's Current Performance

It was hoped that Navoi LC would become the Central Asian Hub for air cargo transportation. However, did the center have all the required characteristics of a Hub? From the point of view of its Resources and Capabilities it had such resources as, modern technology from South Korea, two Boeing-767-300BCF freighters, large land, and strategically valuable location, which granted it such capabilities as, the ability to handle 300 tons of cargo per day and store up to 1000 tons of cargo in any weather, do 40 flights per week and offer a wide range of logistics services. (Berdiyev, 2017, Uzbekistan Airways, 2018) The center's performance was great in the beginning. From 2009 to about 2016 the cargo inflow rose from 18 to 37 thousand tons per year, but then the figures dropped. In 2017 it was only 33 thousand tons. Hence, over the 8 years the center handled only 300 thousand tons of cargo, which is less than 20% of its capacity. And in 2016 it was estimated that the center only used 15% of its total capacity. (Uzbekistan Airways, 2016) It is obvious that the LC's capabilities were highly underutilized.

2.3 Reasons for Navoi LC's failure

Underutilization was a critical reason for the Navoi LC's failure. There exist a number of different opinions that explain the cause of this fatal issue. The President of Uzbekistan, Shavkat Mirziyoyev, and government officials state that the management of the centre was not skilled and dedicated enough to implement effective strategies and undertake proper decision making. Also, they claim that the high freight tariffs were too high and not attractive for foreign airlines. Constantly poor fuel supply was stifling airport efficiency. No coordination between employees was limiting the multimodal capabilities of the LC and the centre's infrastructure was adequate enough to meet global standards (Abdullayev, 2017). However, in addition to the above, the management failed to design an effective digital strategy and implement advanced analytical tools into the decision-making. In this case, Big data solutions can offer broad opportunities to the LC to improve internal efficiency and global competitive advantage. Below, we review some of the Big data solutions for logistics.

3 Big Data Solutions for Logistics

Big data analytics relates to a large-scale data collection, management, and processing characterized by massive data size (volume), rapid data inflow (velocity), and diversity of format and context (variety) (Laney 2001; Yan et al., 2019). Significance of Big data in developing countries is even greater. Big data can accumulate and provide digital assets, i.e. data bundles that represent economic or social value, that can be used by the government, business or citizens for problem solving (Yan et al. 2019; Walker, 2014). Logistic activities yield massive amounts of data and, hence, logistics can serve as a capable data source for Big data initiatives (Nguyen et al., 2017). Big data analytics can provide logistic centers, such as Navoi LC, with many opportunities ranging from improving the efficiency of internal activities to achieving competitive advantage in the global market. Specifically, Big data analytics can help logistics centers improve customer experience (Tan, 2015). DHL's cloud-based big data initiative can serve as a good example (Fig 2). Its agent vehicle tracking system enabled the company to increase efficiency and service quality (Dan et al., 2019). In similar vein, Big data solutions can be implemented at Navoi LC to improve the following: first, to improve the efficiency (lower cost, reduce emissions, etc.), second, to optimize the route design for both internal as well as for external routes, and the third, to utilize the advanced data collection and insight to introduce new services.

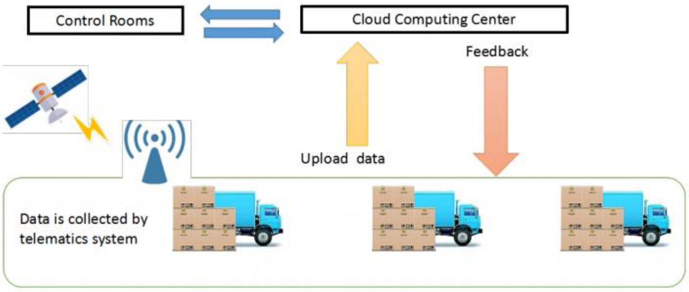


Fig.2. DHL's agent vehicle tracking system (Dang et al., 2019)

4 More Reasons of Failure

Based on the research undertaken by us, it was identified that the center might have suffered because of the change in management. Initially, the majority of the personnel at the LC were professionals from South Korea, but later on their proportion shrined in respect to Uzbek employees, which are less skilled and experienced. It was also found that the Navoi LC had very low quality non-aeronautical services, which are non-technical services created for tourists and business people. They include duty-free shops, logistics parks, business lounges, and other attractive facilities. For all big airports they play an integral role in boosting their popularity and attracting more businessmen. Lack of such amenities would mean a negative impact on the popularity of the airport because of low quality services and overall unattractiveness.

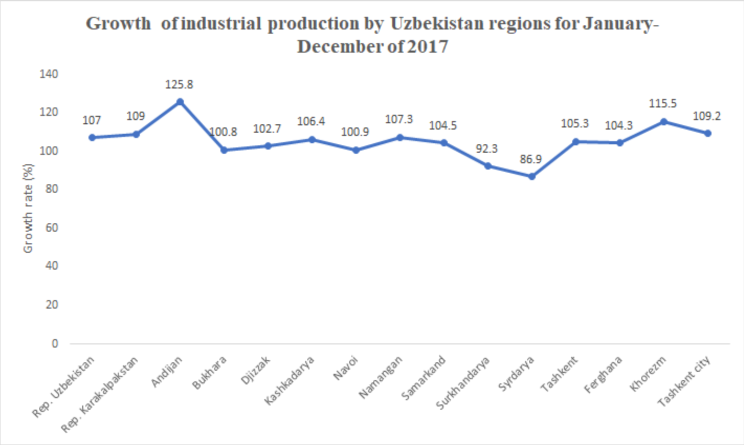


Fig. 3. Growth of industrial production in Uzbekistan (Sharipov, 2017)

Another issue is linked to the recent monopolistic state of Uzbekistan's air industry. Monopolies are notorious for causing inefficiency in the domestic market and a headache for the government due to infant industries. Uzbekistan has to slowly eliminate these inefficient practices one at a time. However, until that is done the air industry will continue suffer. Uzbekistan has issues with meeting production standards. This is a common problem for every developing country because accepting standards takes time and Uzbekistan has started very recently.

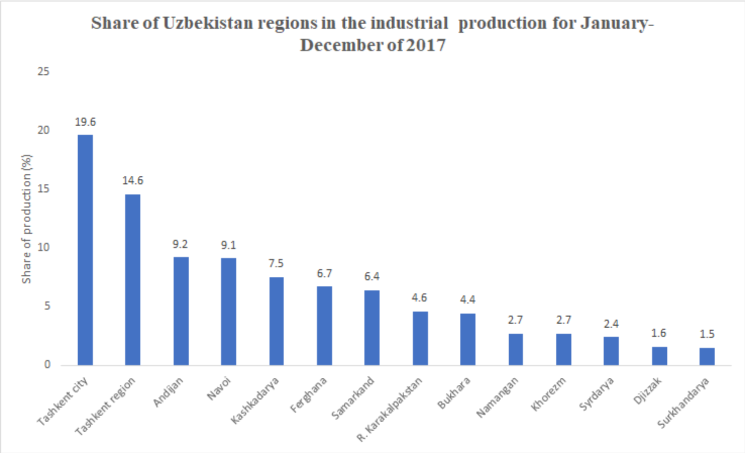


Fig. 4. Share of industrial output in Uzbekistan (Sharipov, 2017)

A major problem that was identified through this research was the highly underdeveloped Hinterland of Navoi. The quality of the hinterland is a key factor for the success of an LC that wishes to dedicate a major part of its operations on exporting. Unfortunately, Navoi LC wished to boost exports but could not manage due to poor hinterland. The gov believed that the region's output would boost the country's exports. But 2016-2017 statistics show that the region with the highest industrial output is not Navoi but Tashkent, and the region with the highest industrial growth is Andijon, not Navoi (Fig 3, 4). In terms of the number of farm enterprises, Navoi region had one of the lowest indicators, while Samarkand had the highest (Sharipov, 2017).

From the consumer market size point of view, the perimeter around Navoi has only 1 million people, and Tashkent has more than twice that amount, which is a bad sign for Navoi's hinterland (Fig 5). And it is also important to mention that Tashkent is closer to the Chinese market than Navoi. That's why Navoi LC had cargo insufficiency. The LC could handle 300 tons per day with the wish to increase it to 1000 tons, but was only receiving 100 tons. (Hudoynazarova, 2015) This has led to a vicious loop, because if there's no cargo, there are no airlines. And if there are no airlines, there is no cargo. Thus, without high value domestic cargo, Navoi will have difficulties attracting many airlines. Hence, Navoi's hinterland requires urgent attention.

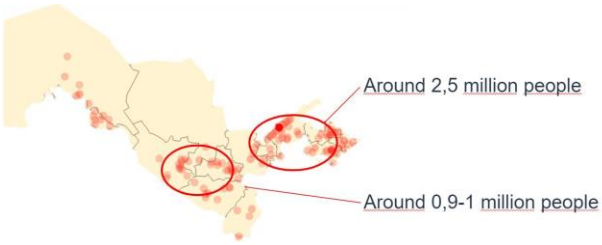


Fig. 5. Population density in Uzbekistan

5 Conclusion

Big Data is gaining increasing relevance due to the globalization trend and technological advances happening worldwide. Truly, in an age where every action can be translated into data, collected, stored, and used for better decision making the use of data analytics is becoming a standard for government institutions and firms. Bid Data has the capability to yield sustainable competitive advantage from accumulated data and assist businesses eliminate inefficiency though improved decision making. (Chen et al., 2012, Muhtaroglu et al., 2013) This is particularly true for modern logistics. Given the fact that logistics systems are constantly intertwined with other sciences and are becoming more complex due to continuous market expansions, Big Data analysis can help supply chains be more visible, flexible, lean, and well integrated. (Genpact, 2014) In the case on the Navoi LC, it is evident that the management lacked in proper strategic decision making, data exchange between transportation modes and departments was poor, and supply of vital resources, such as airplane fuel, was poorly scheduled. We believe that with the help of Big Data analytics the stated problems could have been avoided and the LC would not have suffered from inefficient operations and inappropriate management.

In order to change the situation for the better, the Uzbek government started a package delivery system in at the Navoi LC in 2018 to increase cargo handling and meet the increasing demand for online shopping. In addition, e-documentation was implemented to ease customs procedures for businessmen. The recent meeting with IATA has initiated reform in the domestic air industry, such as tariff changes, to meet international standards. Immense reforms of the inefficient customs procedures were undertaken by the government to ease exporting and importing procedures and attract foreign businesses. To better control and develop the transportation industry in Uzbekistan the Ministry of transport was created, which can help the country develop its multimodal transportation abilities. And the anti-corruption committee was established with the intention to eliminate dishonest and underperforming managers.

All of these changes have solved many problems related to logistics but could not save the Navoi LC. If we look at the requirements for becoming a Hub, which is what the LC was intending to do, it is clear that the government had to still work with Navoi's poorly developed hinterland, the nation's lack of skilled labor, and inefficient Logistics Service Providers (LSP).

So that is why we have proposed to, first of all, promote the center through the use of social media and other types of marketing, have more high quality non-aeronautical services. Attract more Low Cost Carries (LCCs) for better facility utilization, increase in efficiency, cargo and pax. Help the local airline develop by letting it join alliances. And, most importantly, develop the hinterland and retrain local staff with the help of foreign specialists.

References

1. Abdullayev, A.: President found the Logistics Center "Navoi" Ineffective. Gazeta.uz. Retrieved from: https://www.gazeta.uz/ru/2017/12/07/navoi/ (2017)

2. Berdiyev, X.: Logistics Center "Navoi" increased cargo by 8% this year. Podrobno.uz. Retrieved from: https://podrobno.uz/cat/economic/logisticheskiy-khab-navoi-v-etom-godu-uvelichil-perevozki-na-8/ (2017)

3. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q., 4 (36) , pp. 1165--1188 (2012)

4. Dang, S., Shi, J., Li, Y.: Big Data Management in Transport & Logistics Industry: A Literature Review. Journal of Business School, 2(3), 56--62 (2019)

5. Fernie J, Sparks, L.: Logistics and retail management: emerging issues and new challenges in the retail supply chain. Kogan Page (2014)

6. Genpact: Supply chain analytics. Retrieved from: http://www.genpact.com/docs/resource-/supply-chain-analytics (2014)

7. Hudoynazarova, H.: International Logistics Center “Navoi”. Biznes-Daily. Retrieved from: http://www.biznes-daily.uz/uz/gazeta-birja/29159-mjdunarodniy-logistichskiy-sntr-lnavoir)-2015 (2015)

8. Kovács, GY., Kot, S.: New Logistics and Production Trends as the Effect of Global Economy Changes. Polish Journal of Management Studies. vol. 14, No.2 (2016)

9. Laney, D.: 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70), (2001)

10. Muhtaroglu, F.C.P., Demir, S., Obali, M., Girgin, C.: Business model canvas perspective on big data applications. In: Proceedings of the IEEE International Conference on Big Data, pp. 32–-36 (2013)

11. Nguyen, T., Zhou, L., Spiegler, V.: Big data analytics in supply chain management: A state-of-the-art literature review. Computers and Operations Research, 7, 254—265 (2017)

12. Sharipov, J.: Development of industrial production in the Republic of Uzbekistan for January-December 2017. The State Committee of the Republic of Uzbekistan on Statistics. Retrieved from: https://www.stat.uz/uploads/docs/Sanoat-eng-12-2017.pdf (2017)

13. Statista. Forecast for global retail sales growth 2008-2018 | Statista. (2019). Retrieved 19 August 2019, from https://www.statista.com/statistics/232347/forecast-of-global-retail-sales-growth/

14. Tan, K., Zhan, Y., Ji, G.: Harvesting Big Data to enhance supply chain innovation capabilities: an analytic infrastructure based on deduction graph. International Journal of Production Economics, 223—233 (2015)

15. Uzbekistan Airways.: International airport "Navoi". Uzbekistan Airways. Retrieved from: https://www.uzairways.com/en/flights/international-airport-navoi (2016)

16. Uzbekistan Airways.: “Navoi” regional airport is developing logistical infrastructure. Uzbekistan Airways. Retrieved from: https://www.uzairways.com/ru/news/aeroport-navoi-razvivaya-logisticheskuyu-infrastrukturu?field_news_date_value[value]=&page=91 (2018)

17. Walker, J. S.: Big data: a revolution that will transform how we live, work, and think. Int J Advert 33:181--183 (2014)

18. Yan, Z., Ismail, H., Chen, L., Zhao, X., Wang, L.: The application of big data analytics in optimizing logistics: a developmental perspective review. Journal of Data, Information and Management, 1--11 (2019).

Data-Driven Multiplicative Fault Detection Using Hybrid of Multivariate Statistical Techniques

Jungwon Yu, Jinhong Kim, Youngjae Lee, Soyoung Yang and Kil-Taek Lim

Electronics and Telecommunications Research Institute, Daegu-Gyeongbuk Research Center,
1 Techno Sunwan-ro 10gil, Yuga-eup, Dalseong-gun, Daegu, 42994, South Korea
{gardenyoo, jinhong, lyj4295, syyang, ktl}@etri.re.kr

Abstract. Accurate and timely detection of possible faults is indispensable for safe and cost-effective operations of industrial plants. In this paper, we present the results of applying an integration of auto-associative kernel regression (AAKR) and dynamic independent component analysis (which is proposed by Yu et al. [1]) to multiplicative fault detection (FD); in this method, after extracting several latent variables (i.e., independent components) from residual vectors generated by AAKR, detection indices are calculated based on the latent variables and FD is then performed via statistical hypothesis tests. The FD performance of the hybrid method is evaluated with a benchmark example relevant with multiplicative fault type, and compared with various popular FD methods. The experimental results show that the hybrid method achieves the lowest type II error (i.e., miss detection rate) and, at the same time, acceptable type I error (i.e., false alarm rate).

Keywords: multiplicative fault, fault detection, auto-associative kernel regression, dynamic independent component analysis

1 Introduction

In modern industrial processes, such as power plants, and manufacturing and chemical processes, accurate and timely detection of potential faults can improve availability, safety, and reliability of them, and permit their cost-effective operations; here, the faults are defined as abnormal events occurring in operating processes. With the recent development of sensor, measurement, database, and communication technologies, the quality and quantity of the data gathered from the target processes have been raised dramatically; the massive amounts of multivariate data composed of various process variables can be collected and managed efficiently. Therefore, much attention will be focused on minimizing downtime and maximizing process efficiency via analyzing the multivariate process data.

The types of the faults can be classified by examining whether mean values of process variables have changed or whether variance-covariance structures of them have changed [2-4]; the former and the later are referred to as additive and multiplicative faults, respectively. Additive fault can be further divided into abrupt and incipient faults. Abrupt fault corresponds to the faults in which mean vectors of process variables have changed abruptly; there are several examples of abrupt fault, such as biased sensor measurement and tube ruptures. Incipient fault is the faults in which the mean vectors slowly change; here, their deviations from normal operating regions become larger gradually. Drift sensor faults, tube leaks, and slow degradation of components are typical examples of incipient fault. Contrary to additive fault in which the mean vectors change over time, multiplicative fault affects variance-covariance structures of process variables. Although multiplicative faults may not lead to system failures directly, they are closely related with lifetime of target processes.

Multiplicative faults with abnormal variance-covariance structures of process variables can be formulated as [2]

$$\mathbf{x}(k) = \mathbf{F}\mathbf{x}^{*}(k) \tag{1}$$

where the data vectors $\mathbf{x}(k)$ and $\mathbf{x}^{*}(k)$ are contaminated vector by the fault effects and normal vector without the effects at time k, respectively, and $\mathbf{F}$ is a faulty gain matrix; here, if target systems are normal, $\mathbf{F} = \mathbf{I}$; otherwise, $\mathbf{F} \neq \mathbf{I}$. Multiplicative faults described in eq. (1) not influence the mean vector of $\mathbf{x}(k)$, but increase the variations of target process variables.

In this paper, we present the results of applying a hybrid of auto-associative kernel regression (AAKR) and dynamic independent component analysis (DICA) (which is proposed by Yu et al. [1]) to multiplicative fault detection (FD); this paper is intended to verify the performance of the hybrid method for multiplicative FD, which is an extension of the previous work reported by Yu et al. [1]. It should be noted that studies on data-driven FD methods for multiplicative faults seem to be lacking compared with those for additive faults. In the method (hereafter, indicated by 'AAKR+DICA'), residual vectors are generated by AAKR, and residual analysis via DICA is carried out to capture useful information hidden in the residuals. AAKR is a nonparametric multivariate technique to generate predicted vectors corresponding to new query vectors by real-time local modeling; one does not need to be concerned about the properties of target data (i.e., 'linear or nonlinear' and 'unimodal or multimodal') in advance. In standard AAKR, FD can be performed based on squared prediction error (SPE) obtained from the residual vectors (i.e., difference between actual query vectors and predicted vectors by AAKR). The problem is that SPE-based FD cannot take into account the following facts. First, each component of the residual vectors may not follow normal distribution exactly. Second, the residual components may be correlated with each other and there may exist serial-correlations in each component. It is worthwhile to emphasize that residual analysis based on DICA is capable of considering not only the non-Gaussianity but also the statistical relationships (i.e., cross- and auto-correlations). In 'AAKR+DICA', detection indices for FD are calculated from several latent variables extracted by DICA.

To verify the multiplicative FD performance, the hybrid method (i.e., 'AAKR+DICA') and several comparison methods, such as local outlier factor (LOF), principal component analysis (PCA), standard ICA and AAKR, are applied to benchmark problem [2] relevant with multiplicative fault type. The experimental results demonstrate that the hybrid method can successfully detect the multiplicative fault; the method achieves the lowest type II error (i.e., miss detection rate) and, at the same time, acceptable type I error (i.e., false alarm rate) lower than significance level $\alpha = 0.01$.

The remainder of this paper is organized as follows. Section 2 briefly summarizes the integration method (i.e., 'AAKR+DICA') proposed by Yu et al. [1]. In Section 3, after explaining the benchmark problem associated with multiplicative fault introduced in [2], we present the results of applying 'AAKR+DICA' and the comparison methods to the benchmark problem. Finally, we give our conclusions in Section 4.

2 Summary of the hybrid method

In this section, we briefly summarize the hybrid method ('AAKR+DICA') proposed by Yu et al. [1]; for more details, readers are invited to read the Refs. [1, 4]. Detailed descriptions of AAKR and DICA can be found in the Refs. [5-8] and [9-14], respectively.

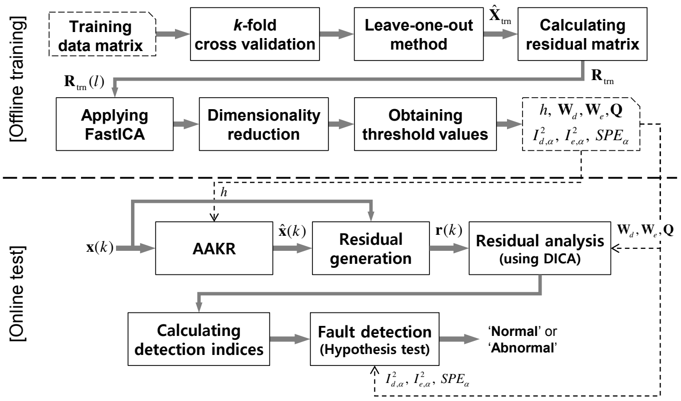

[Offline training]
Training data matrix
k-fold cross validation
Leave-one-out method
$\hat{\mathbf{X}}_{\text{trn}}$
Calculating residual matrix
$\mathbf{R}_{\text{trn}}$
$\mathbf{R}_{\text{trn}}(l)$
Applying FastICA
Dimensionality reduction
Obtaining threshold values
h, $\mathbf{W}_d, \mathbf{W}_e, \mathbf{Q}$
$I^2_{d,\alpha}$, $I^2_{e,\alpha}$, SPE_α
[Online test]
h
$\mathbf{x}(k)$
AAKR
$\hat{\mathbf{x}}(k)$
Residual generation
$\mathbf{r}(k)$
Residual analysis (using DICA)
$\mathbf{W}_d, \mathbf{W}_e, \mathbf{Q}$
Calculating detection indices
Fault detection (Hypothesis test)
'Normal' or 'Abnormal'
$I^2_{d,\alpha}$, $I^2_{e,\alpha}$, SPE_α

Fig. 1. Schematic diagram of the hybrid method [1].

The FD procedure based on 'AAKR+DICA' is described in Fig. 1 [1]; this procedure is roughly divided into offline training and online test phases. In the offline training phase, first, to determine the kernel parameters (e.g., bandwidth parameter h of Gaussian function), k-fold cross validation is applied to the training data matrix $\mathbf{X}_{\text{trn}}$. Second, after calculating the predicted data matrix $\hat{\mathbf{X}}_{\text{trn}}$ using the leave-one-out method, residual matrix $\mathbf{R}_{\text{trn}}$ (= $\mathbf{X}_{\text{trn}} - \hat{\mathbf{X}}_{\text{trn}}$) is obtained. Third, the matrix $\mathbf{R}_{\text{trn}}$ is extended into the augmented matrix $\mathbf{R}_{\text{trn}}(l)$, and fastICA algorithm [12-14] is then used to estimate the matrices $\mathbf{W}$, $\mathbf{B}$, and $\mathbf{Q}$ from the matrix $\mathbf{R}_{\text{trn}}(l)$ (in general, the number of lags l is set to 1 or 2); by using the dimensionality reduction, the matrices $\mathbf{W}$ and $\mathbf{B}$ can be partitioned into dominant parts $\mathbf{W}_d$ and $\mathbf{B}_d$, and excluded parts $\mathbf{W}_e$ and $\mathbf{B}_e$, respectively. Finally, detection indices, I_d^2, I_e^2, and SPE statistics, are calculated from the training samples, and their threshold values, $I_{d,\alpha}^2$, $I_{e,\alpha}^2$, and SPE_α, are predetermined through kernel density estimation.

In the online test phase, after obtaining the predicted vector $\hat{\mathbf{x}}_{\text{test}}(k)$ for new query vector $\mathbf{x}_{\text{test}}(k)$ at time k , residual vector $\mathbf{r}_{\text{test}}(k)$ (= $\mathbf{x}_{\text{test}}(k) - \hat{\mathbf{x}}_{\text{test}}(k)$) is generated, and it is also extended into the augmented vector with l lagged vectors. To determine whether target systems are normal or faulty, statistical hypothesis testing should be performed; here, the detection indices are calculated, and if one of them is larger than the predefined limits, occurrences of faults are declared and alarm signals are generated. Generally, the alarm signals without consistency are ignored and regarded as false alarms; when they are continuously observed, faulty variables should be isolated and root causes of occurred faults should be carefully investigated.

3 Experiment results

In this section, 'AAKR+DICA' and comparison methods (such as LOF, PCA, ICA and AAKR) are applied to benchmark data [2] related with multiplicative fault type; this section is based on dissertation research completed in [4]. Simulation data is generated by the following open-loop synthetic linear equation:

$$\begin{aligned} \mathbf{y}(k) &= \mathbf{W}\mathbf{x}(k) + \boldsymbol{\xi}(k) \\ \boldsymbol{\theta}(k) &= \boldsymbol{\Psi}\mathbf{y}(k) + \mathbf{b} + \boldsymbol{\eta}(k) \end{aligned} , \tag{2}$$

where $\mathbf{x}(k) \in \Re^3$, $\mathbf{y}(k) \in \Re^{15}$, $\boldsymbol{\theta}(k) \in \Re^2$, $\mathbf{W}$ = rand(15, 3) $\in \Re^{15\times 3}$, $\boldsymbol{\Psi}$ = 1+rand(2, 15) $\in \Re^{2\times 15}$, $\mathbf{b} = [1055\ 1825]^T$, and $\mathbf{x}$, $\boldsymbol{\xi}$, and $\boldsymbol{\eta}$ follow multivariate normal distributions, i.e., $\mathbf{x} \sim MVN_3(\mathbf{0}, \text{diag}(4, 1.96, 1.44))$, $\boldsymbol{\xi} \sim MVN_{15}(\mathbf{0}, 10^{-6}\mathbf{I}_{15})$, and $\boldsymbol{\eta} \sim MVN_2(\mathbf{0}, 10^{-8}\mathbf{I}_2)$, where $\mathbf{I}_{15} \in \Re^{15\times 15}$ and $\mathbf{I}_2 \in \Re^{2\times 2}$ are identity matrices; each component of rand$(\cdot,\cdot)$ is a uniform random number between 0 and 1. To train FD models, 500 normal data vectors are generated based on eq. (2); each data vector is composed 17 process variables, i.e., $[\mathbf{y}^T(k)\ \ \boldsymbol{\theta}^T(k)]^T$. To verify the FD performance, two cases of multiplicative faulty dataset are prepared; after generating 500 normal samples by eq. (2), the following two multiplicative fault effects are injected from time k = 101 to the end:

- Case 1: The entry in the 2nd row and 1st column of the matrix $\boldsymbol{\Psi}$ becomes bigger 10 times, i.e., $\boldsymbol{\Psi}(2, 1) = 10\times\boldsymbol{\Psi}^*(2, 1)$, from time k = 101 to the end.
- Case 2: From time k = 1 to 100, faulty gain matrix $\mathbf{F}$ is equal to identity matrix $\mathbf{I}_{15} \in \Re^{15\times 15}$, i.e., $\mathbf{y}(k) = \mathbf{F}\mathbf{y}^*(k)$ where $\mathbf{F} = \mathbf{I}_{15}$; but if a multiplicative fault has occurred (from time k = 101 to the end), the number 5 is added into 3rd row and 6th column of $\mathbf{F}$, i.e., $\mathbf{F}(3, 6) = \mathbf{F}^*(3, 6) + 5$.

In the above descriptions, superscript '*' indicates that the corresponding parameter matrices are normal.

First, let us take a look at the distribution shapes of residuals generated by leave-one-out method. Fig. 2 shows histogram and normal probability plots of several residuals, i.e., r_1^i, r_2^i, and r_3^i (i = 1,..., 500), among 17 residual components; the plots for the others are similar to those of r_1^i, r_2^i, and r_3^i, so they are omitted to save space. Normal probability plot is a graphical tool to confirm how closely the target data follow Gaussian distribution; the closer the data distribution is to normal, the nearer the positions of data points (indicated by '+') are to the dash-dot red lines. As can be seen from the figure, both tails are asymmetric each other, and deviate from Gaussian distribution significantly. Therefore, it is possible to enhance the performance of FD by applying ICA to the non-Gaussian residuals.

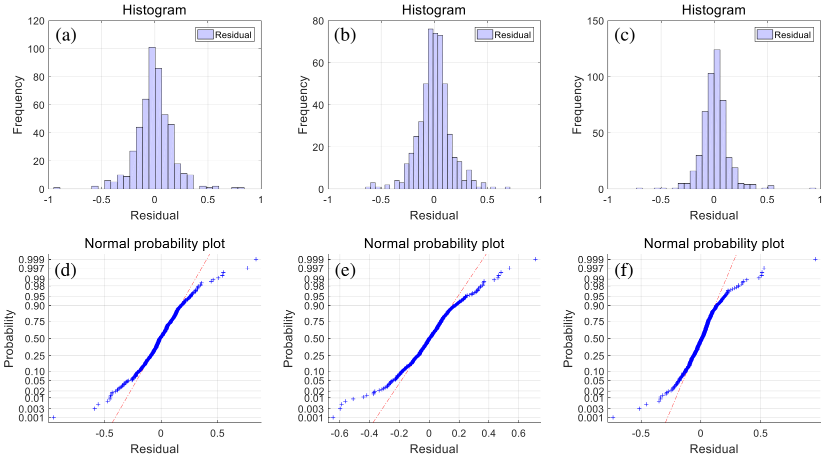


Fig. 2. Histogram and normal probability plots of the residuals (r_1^i, r_2^i, and r_3^i) obtained from the 500 training samples: (a) histogram of r_1^i; (b) histogram of r_2^i; (c) histogram of r_3^i; (d) normal probability plot of r_1^i; (e) normal probability plot of r_2^i; (f) normal probability plot of r_3^i [4].

Table 1 lists the experimental results of applying 'AAKR+DICA' and the comparison methods to the two multiplicative fault cases. The type I and II errors are obtained by averaging 100 independent simulations under the same conditions. In Case 1, type II errors of LOF, PCA, and AAKR are larger than 95%; these methods cannot detect the multiplicative fault in Case 1 properly. Type II errors of ICA and 'AAKR+DICA' are lower than 10%, which are suitable for being applied to Case 1; false alarm rates of them are nearly the same, but 'AAKR+DICA' improves FD rate by about 8% compared with ICA. In Case 2, among the comparison methods, PCA and ICA shows the highest and the lowest type II errors, respectively; type II error of 'AAKR+DICA' is slightly lower than ICA. In the table, comparing the errors of AAKR and 'AAKR+DICA' proves clearly that residual analysis via DICA is powerful tool for FD. Figs. 3 and 4 show the monitoring charts whose type II errors are closest to those of the averaged in Table 1 among the 100 times experiments; *Y*-axes of the figures for ICA, AAKR, and 'AAKR+DICA' are indicated in logarithmic scale. As shown in the figures, 'AAKR+DICA' exhibits superior performance compared with the others. The evidence for the effectiveness of DICA-based residual analysis can be seen in Figs. 3 (d) and (e) and Figs. 4 (d) and (e), respectively.

Table 1. Performance indices of the proposed and comparison methods (lower is better) [4].

	LOF		PCA Q statistic		ICA I_e^2		AAKR SPE		‘AAKR+DICA’ I_d^2	
	type I	type II	type I	type II	type I	type II	type I	type II	type I	type II
Case 1	1.14	95.83	0.41	99.35	0.51	8.48	0.75	95.99	0.43	0.04
Case 2	1.07	23.01	0.39	50.16	0.48	0.08	0.94	33.58	0.45	0.00

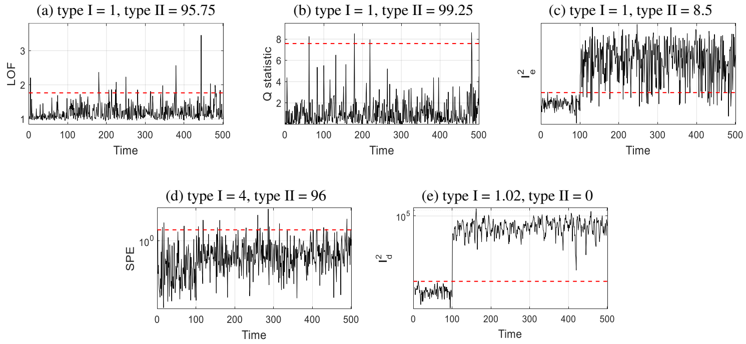


Fig. 3. (Case 1) Detection indices for 500 test data samples of the comparison and proposed methods: (a) LOF value; (b) Q statistic of PCA; (c) I_e^2 of ICA; (d) SPE of AAKR; (e) I_d^2 of ‘AAKR+DICA’ [4].

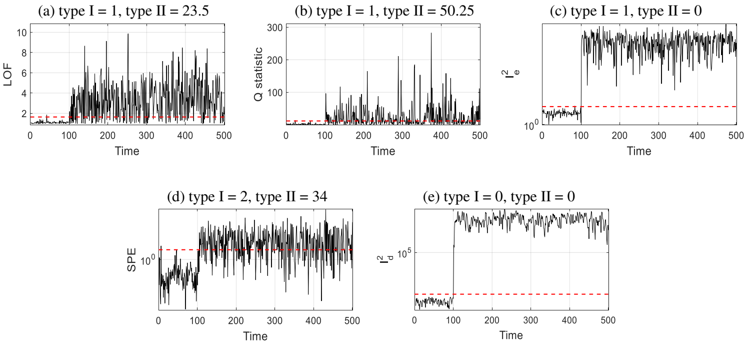


Fig. 4. (Case 2) Detection indices for 500 test data samples of the comparison and proposed methods: (a) LOF value; (b) Q statistic of PCA; (c) I_e^2 of ICA; (d) SPE of AAKR; (e) I_d^2 of ‘AAKR+DICA’ [4].

Next, let us look at statistical relationships between generated residuals by AAKR. Fig. 5 shows the correlation coefficients between 17 residuals in the normal region (from time k = 1 to 100) for Case 1; as the coefficients are close to +1 (or −1), the colors for relevant entries are close to black (or white). In this figure, it is confirmed that residuals are highly correlated to each other; this can be attributed to the fact that, as described in eq. (2), original monitoring variables are highly correlated to each other. Fig. 6 shows the scatter plots of some residual pairs (i.e., $r_1(k)$ vs. $r_2(k)$, $r_2(k)$ vs. $r_7(k)$, and $r_3(k)$ vs. $r_{11}(k)$) in the normal region (k = 1,..., 100); correlation coefficients between X- and Y-axes variables are given in the bottom left or right of each figure. In this figure, we can confirm that there exist clear positive or negative correlations in the pairs. In the ‘AAKR+DICA’, these statistical characteristics of residuals can be properly handled via DICA; therefore, it achieves better FD performance than standard AAKR.

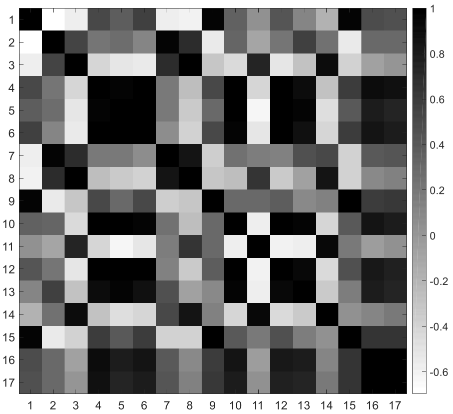


Fig. 5. Correlation coefficients between 17 residuals generated by applying standard AAKR to the normal region (time k = 1 to 100) of Case 1 [4].

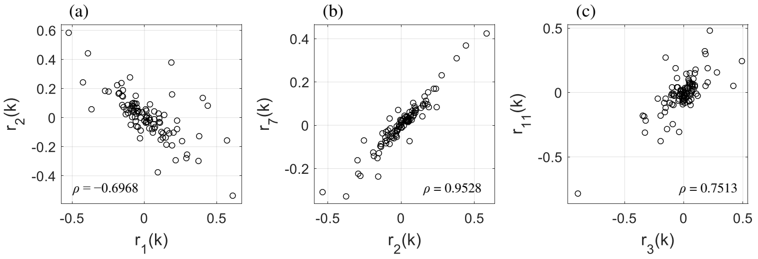


Fig. 6. Scatter plots between some residual pairs associated with normal test samples (k = 1,..., 100) of Case 1: (a) $r_1(k)$ vs. $r_2(k)$; (b) $r_2(k)$ vs. $r_7(k)$; (c) $r_3(k)$ vs. $r_{11}(k)$ [4].

4 Conclusion

Over the past few decades, a considerable number of studies have been made on FD methods for additive faults; what seems to be lacking is the study for data-driven multiplicative FD. In this paper, we employed the integration method (i.e., ‘AAKR+DICA’) proposed by Yu et al. [1] to detect multiplicative faults; this method and several comparison methods, such as LOF, PCA, standard ICA and AAKR, were applied to the benchmark example [2] associated with multiplicative fault type. The experimental results demonstrated that ‘AAKR+DICA’ can detect the multiplicative fault cases successfully; the method exhibited superior FD performance (i.e., the lowest type II error) compare to the comparison methods. AAKR has the advantage that it can generate residual vectors in light of the nonlinearity and multimodality of the target processes; furthermore, residual analysis via DICA can reveal the underlying fundamental structure hidden in the residuals. Therefore, the integration can create significant synergies in FD for complex industrial processes.

Acknowledgments. This material is based upon work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under i-Ceramic platform construction project (i-Ceramic manufacturing innovation platform). No. 20004367, 'Development of Cloud Big Data Platform for the Innovative Manufacturing in Ceramic Industry'

References

1. J. Yu, J. Yoo, J. Jang, J. H. Park, and S. Kim, "A Novel Hybrid of Auto-Associative Kernel Regression and Dynamic Independent Component Analysis for Fault Detection in Nonlinear Multimode Processes," J. of process control, vol. 68, pp. 129-144, Aug. 2018.
2. H. Hao, K. Zhang, S. X. Ding, Z. Chen, and Y. Lei, "A data-driven multiplicative fault diagnosis approach for automation processes," ISA Trans., vol. 53, no. 5, pp. 1436-1445, Sep. 2014.
3. K. Zhang, Y. A. Shardt, Z. Chen, and K. Peng, "Using the expected detection delay to assess the performance of different multivariate statistical process monitoring methods for multiplicative and drift faults," ISA Trans., vol. 67, pp. 56-66, Mar. 2017.
4. J. Yu, "Data-Driven Fault Detection and Isolation of Complex Industrial Processes Using Hybrid of Multivariate Statistical Techniques," Ph.D. dissertation, Dept. of Elect. and Comput. Eng., Pusan Nat. Univ., Busan, South Korea, 2018.
5. J. W. Hines, D. Garvey, R. Seibert, A. Usynin, and S. A. Arndt, "Technical Review of On-Line Monitoring Techniques for Performance Assessment (NUREG/CR-6895) Vol. 2, Theoretical Issues," Published May, 2008.
6. J. Garvey, D. Garvey, R. Seibert, and J. W. Hines, "Validation of on-line monitoring techniques to nuclear plant data," Nucl. Eng. and Technol., vol. 39, no. 2, pp. 149-158, Apr. 2007.
7. J. W. Hines, and D. R. Garvey, "Development and application of fault detectability performance metrics for instrument calibration verification and anomaly detection," J. of Pattern Recognition Res., vol. 1, no. 1, pp. 2-15, 2006.
8. J. Yu, J. Jang, J. Yoo, J. H. Park, and S. Kim, "Bagged auto-associative kernel regression-based fault detection and identification approach for steam boilers in thermal power plants," J. of Elect. Eng. & Technol., vol. 12, no. 4, pp. 1406-1416, 2017.
9. J. M. Lee, C. Yoo, and I. B. Lee, "Statistical process monitoring with independent component analysis," J. of Process Control, vol. 14, no. 5, pp. 467-485, Aug. 2004.
10. J. Lee, B. Kang, and S. H. Kang, "Integrating independent component analysis and local outlier factor for plant-wide process monitoring," J. of Process Control, vol. 21, no. 7, pp. 1011-1021, Aug. 2011.
11. J. M. Lee, C. Yoo, and I. B. Lee, "Statistical monitoring of dynamic processes based on dynamic independent component analysis," Chem. Eng. Sci., vol. 59, no. 14, pp. 2995-3006, Jul. 2004.
12. A. Hyvärinen, and E. Oja, "Independent component analysis: algorithms and applications," Neural Netw., vol. 13, no. 4-5, pp. 411-430, June 2000.
13. A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
14. A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," IEEE Trans. Neural Netw., vol. 10, no. 3, pp. 626-634, May 1999.

Machine Nonactive State Detection using Statistical Features and K-Nearest Neighbor

Taing Borith[1], In Joo[1], Aziz Nasridinov[1] , Kwan Hee Yoo[1*]

[1]Chungbuk National University, South Korea,
*Corresponding Author

Abstract. In the current manufacturing field, an abnormality of machines which we defined as the nonactive state of machines during operating is a crucial issue in the manufacturing area. Therefore, an accurate nonactive state detection model for the factory's machine can lead to improving productivity, avoid catastrophic failures and minimize economic losses, which is a mainstream solution in the manufacturing area. To better detect the fragilities, in this work, we present a novel two-step data-driven for nonactive detection of the industry machine. The feature extraction steps include multiple statistical algorithms which consist of reliability features, proposed feature extraction method, time-domain features, and frequency domain features to extract the active and nonactive patterns from the raw data. In the nonactive state detection step, we constructed a machine learning model by utilizing the K-Nearest Neighbors algorithm to detect an active and nonactive status of industrial machines.

Keywords: Statistical Algorithms, Features Extraction, Machine Learning, KNN

1 Introduction

In modern industry these days, many factories employed continuous flow processes method used in manufacture to produce the materials and maintain the process without interruption. Apart from that, these massive industrial processes also cause great harm to production machines. Along with the tremendous harmful to factory's machines, many factories face significant loss due to the impacts of the unstable operation of the machine. We defined these unstable states of machines as the nonactive state of machines. At any point in time that nonactive states take place during production machine is operating, it will cause vastly lost to the fabricator. Moreover, when plenty of nonactive states occurred during machines operating, they could make machines working slower than the usual process, or sometimes cause machines to produce defective products. In recent year, there are diversities of detection approaches for dealing with the problem in the manufacturing area have been proposed — some approaches using mathematical methods to represent the physics of failure and the phenomenon of degradation. Despite the mathematical approach, some strategies employ deep learning or machine learning to construct the detection model. However, it still faces some difficulties to construct the quality detection model in the manufacturing area. Therefore, to capably deal with these problems, a combination of machine learning with multiple statistical feature extraction is proposed. In this work, by considering of various factors that can cause an unusual state of machine occur during operating, multi-analysis methods such as reliable analysis which include time-between-nonactive, Weibull [1], Log-normal, Exponential Distribution are employed to extract the features of the nonactive pattern of the machine. Additionally, to take the improving productive and detection model accuracy into consideration, we proposed a feature extraction method which can analyze the performance of the machine during operating. Besides that, to better detect active state and nonactive state of the machine, we utilized the time-domain algorithm and frequency-domain algorithm [2] to extract additional features from the machine status tracking value method. Furthermore, we constructed a nonactive state detection of the machine with the K-Nearest Neighbor model [3]. As a result, the nonactive state detection model with KNN got an acceptable score with 93% of accuracy.

1.1 Structure of Propose Method

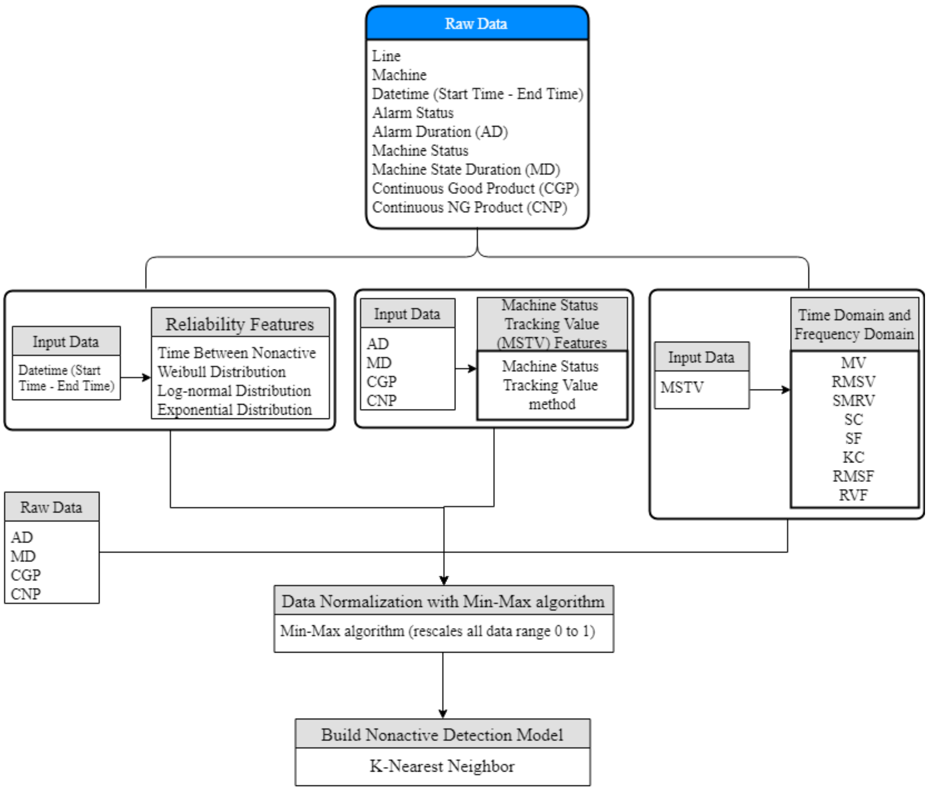

Raw Data
Line
Machine
Datetime (Start Time - End Time)
Alarm Status
Alarm Duration (AD)
Machine Status
Machine State Duration (MD)
Continuous Good Product (CGP)
Continuous NG Product (CNP)
Input Data
Datetime (Start Time - End Time)
Reliability Features
Time Between Nonactive
Weibull Distribution
Log-normal Distribution
Exponential Distribution
Input Data
AD
MD
CGP
CNP
Machine Status Tracking Value (MSTV) Features
Machine Status Tracking Value method
Input Data
MSTV
Time Domain and Frequency Domain
MV
RMSV
SMRV
SC
SF
KC
RMSF
RVF
Raw Data
AD
MD
CGP
CNP
Data Normalization with Min-Max algorithm
Min-Max algorithm (rescales all data range 0 to 1)
Build Nonactive Detection Model
K-Nearest Neighbor

Figure 1. The Global Structure of Machines Nonactive State Detection

As shows in Figure 1 the global structure of machines nonactive state detection. In this work, the data preprocess and feature extract are divided into 3 parts. Firstly, we retrieved raw data from database and calculated the elapsed time between the nonactive state of machines which denote as time-between-nonactive. Then these Time Between Nonactive data is used as life-time data to apply in reliable statistical algorithm to extract reliable features which include Weibull Distribution, Log-normal Distribution, and Exponential Distribution. In the second path, we extract a new feature by combining various important information from raw data set and denote as Machine Status Tracking Value (MSTV). In the third path of feature extract, we used the extracted feature "MSTV" to apply into time-domain features and frequency domain features. Time-domain features include mean value (MV), root mean squirt value (RMSV), Square Mean Root Value (SMRV), Skewness Coefficient (SC), Shape Factor (SF), and Kurtosis Coefficient (KC). Besides that, Root Mean Square Frequency (RMSF) and Root Variance Frequency (RVF) are frequency domain features [2]. In the nonactive state detection step, we combine four useful features from raw data which consist of alarm duration (AD), Machine State Duration (MSD), Continuous Good Product (CGP), and Continuous NG Product (CNP) with all extracted features as input data X for training supervised learning model of KNN algorithm and scale all the features into the range of 0 to 1. Moreover, the raw data named machine state is used as input data Y. Finally, we build the detection model with KNN algorithm which set the K nearest neighbor to 5.

2 Experiment and Result

As mention in the previous section, there are 17 features used as input data X and 1 feature used as input data Y for KNN supervised learning model. Moreover, to speed up the training process and reduce the bias between each feature, we rescale all the features which used as input data X into the range of 0 to 1 using the Min-Max normalization algorithm. In the training model, the k value of the KNN algorithm is set to 5 neighbors. After the model is built, we used independent data set with 86,400 observation and Precision and recall method used as validation matric to validate the accuracy of the model. In Table 2 the Validation score of the nonactive state detection with KNN model, exhibits that the detection model which constructed from KNN algorithm obtain acceptable score with 93% and 95% of Average Precision score and Recall score, respectively.

As a result, the nonactive state detection which is constructed from KNN detected 72,068 of the active state and 14,332 of nonactive state of the machine while the actual nonactive is 13,214 as illustrated in Figure 2 the comparison of actual value and predicted value from KNN model.

Table 2. Validation score of the nonactive state detection with KNN model

Model	Average Precision	Recall	Observation Number
K-Nearest Neighbor	93%	95%	86,400

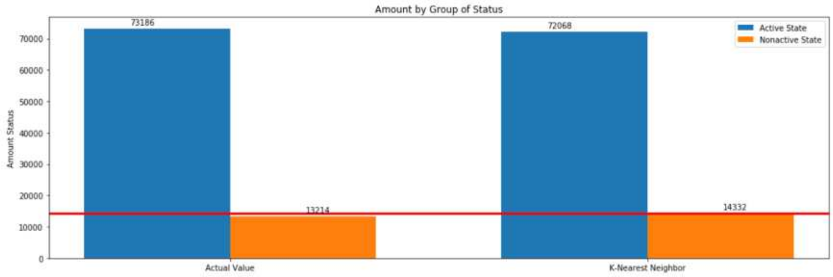


Figure 2. the comparison of actual value and predicted value from KNN model

3 Conclusion

We presented an approach to construct an accurate nonactive state detection of the machine with KNN. The detection model formed by various statistical methods and machine learning-based approach to detect the active state and nonactive state of the machine. This approach utilized reliability analysis, the proposed feature extraction method (MSTV), time-domain and frequency-domain analysis to extract features from the raw data set. According to the experimental results, the predicted outcome from the KNN model achieved 93% of accuracy rate compared to the actual value. It is good result to execute the machine learning to detect active or not of machine in factory using KNN. Due to the statistical methods used to extract the train features is too traditional, and the detection model cannot solve all the issues in the manufacturing fields. Therefore, in the future, we will try to utilize other useful statistical methods such as Hibert-Huang Transform method to enhance the model accuracy.

Acknowledgments. The work is supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under i-Ceramic manufacturing innovation platform technology development business. No.2004367, 'Development of cloud big data platform for the innovative manufacturing in ceramic industry', and by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW (2019-0-01183) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

References

1. M. Xie, C. D. Lai.: Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function. Reliability Engineering & System Safety, ScienceDirect, vol. 52, pp. 87-93, (1996)
2. J. Wu, Ch. Wu, Sh. Cao, S. W. Or, Ch. Deng, X. Shao.: Degradation Data-Driven Time-To-Failure Prognostics Approach for Rolling Element Bearings in Electrical Machines. IEEE Transactions on Industrial Electronics, vol. 66, pp.529-539, Jan. (2019)
3. Zh. Zhou, Ch. Wen, Ch. Yang.: Fault Detection Using Random Projections and k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes. IEEE Transactions on Semiconductor Manufacturing, vol.28, pp.70-79, Feb. (2015)

Data Strategy in Ceramic Science and Industry

Sung Beom Cho and Sangil Hyun

Virtual Engineering Center, Korea Institute of Ceramic Engineering and Technology (KICET), csb@kicet.re.kr

Abstract. Here, we summarized the recent progress in data-driven materials science, in particular ceramic-related field. Since modern academic establishment of ceramic science, the materials and related process has been developed by numerous trials-and-errors. Despite of the guidance of theory and simulation, many of end-product properties are still in the era of the art rather than engineering.

Keywords: Materials Science, Ceramics, Simulation, Materials Genome, Data

Materials engineering is one of the oldest engineering branches. It is because the mankind invented the ceramics and steals in prehistory age. This field is still importantly treated until modern period since the metal, mining, glasses are strongly related to the power of nations. Metallurgy, mining, and glass science is one of the first established department of engineering field in the modern university [1]. However, most of the development and progress in materials, in particular, ceramics, has been established in trials-and-errors manner. Because of infinite chemical space and process parameters, materials discovery and process optimization has been though as the era of art. Even though simulations and theories can guide a confined region, complex nature of materials has not been understood and still remains unpredictable. The prediction of final property of materials is still highly challenging except a handful of simple cases.

Recent advances in data science have led to digitalization and data mining in a diverse science and industry field [2]. However, it has been difficult to exploit the ceramic materials science and industry, especially for materials discovery and process optimization. The data structure of materials is often unstructured and hierarchal, and lots of materials information is hidden and not digitalized. Also, gathering big data is difficult because the cost of each data point set is expensive. Sometimes, it should be prohibited due to physical limitation. Therefore, the materials discovery and process optimization cannot be guided by theoretical prediction and should be done in a case by case manner with a lot of trials and error as known as Edisonian approaches. The complex, hierarchal, and limited data set of materials requires feature engineering and additional data-mining. In this talk, machine learning strategy on materials discovery and processing of ceramic industries will be discussed. For the materials discovery examples, we used first-principles calculations and retrieve the database with filtering technique. With this approach, we can efficiently screen the materials and accelerate the discovery processes. For the process optimization, we picked the tape-casting process and powder-pressing process as a example case. Since the feasibility test on small data set was successful, we discuss the extendable approaches.

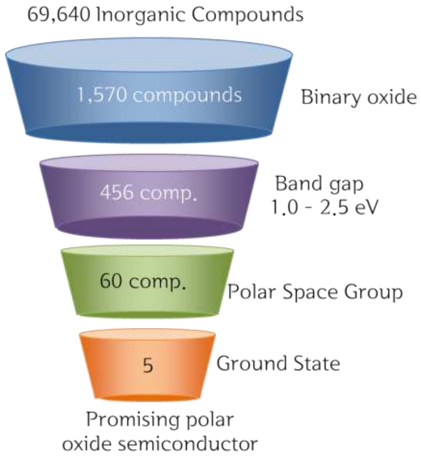


Figure 1. Data Retrieving from First-principle database

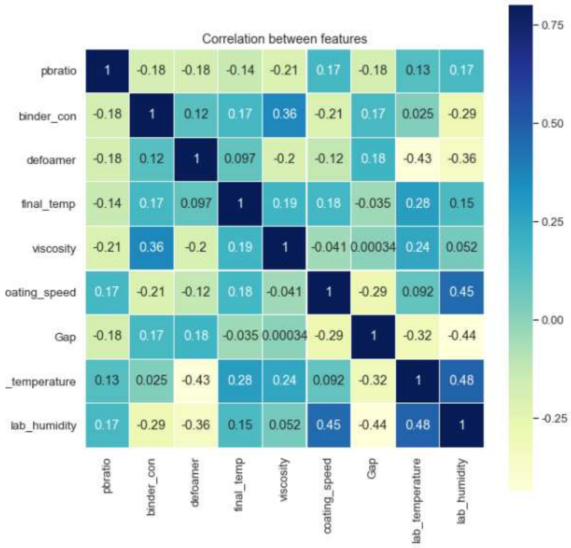


Figure 2. Data correlation of process parameters of tape-casting

References

1. Hummel, Rolf E: Understanding Materials Science History, Properties, Applications. Springer-Verlag New York (2005)
2. Jose, R., Ramakrishna, S., Materials 4.0: Materials big data enabled materials discovery Applied materialstoday 10 pp 127-132 (2018)

A study of Defect Classification in Ceramic Materials using Deep Learning

Hye-Jin S. Kim[1], Suyoung Chi[1],

[1] Electronics and Telecommunications Research Instatitue
{marisan, chisy}@etri.re.kr

Abstract. Ceramic products have been used for a long time. Recently, Ceramic parts have played an important role in smart-phone industries. In this paper, we demonstrate the defect classification technique for determining defects in advanced ceramic materials.

Keywords: ceramic, classification, image, defect

1 Introduction

There have been many attempts to apply artificial intelligence to the industries. Recently, ceramic manufacturing process try adopting not only machine learning but also cloud platform techniques.

Advanced ceramic materials have its own great properties such as high temperature strength; hardness; inertness to chemicals, food; resistance to wear and corrosion. Due to these properties, military and many industrial applications such as smart phone use ceramic parts. However, widespread use is limited due to the presence of defects. Defects are manifested as chipping, voids, and cracks, or inclusions caused by foreign material.

In order to detect surface and subsurface damage in ceramic material, a particular device such low coherence optical scatter reflectometer [1] were adopted. Recently, image processing techniques [2] were used for analysis and detection of surface defects in ceramic materials. In [2], many operations such as image enhancement, edge detection and morphological method were used for classification.

The inapplicable defects in the advanced ceramic material are not able to be detected by the previous methods. Therefore, we introduce CNN-based defect classification technique.

2 Data Preparation and their statistics

The ceramic products are 1,219 in total. We took photos for them in gray scale with bmp format. The image size is 640x480 that corresponds to an object. Our data consist of 5 classes such as normal, chipping, inclusion, processing defect and crack.

Their statistics are presented in Fig. 1. Comparing to other classes, the samples of processing defects and crack are too small to be analyzed. Moreover, most samples in the inclusion class also cover the chipping class. This distribution depends on the ceramic products.

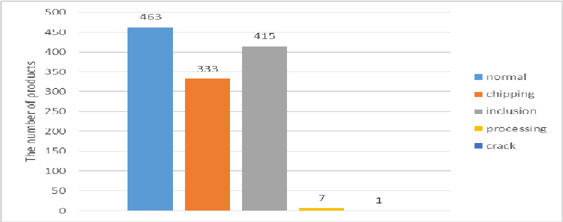


Fig. 1. The statistics of ceramic products.

Fig. 2 and Fig 3. are showed samples of various defect classes and normal class. Because of unevene distribution of data statistics, we divide our data into two classes: defect and normal. Fig. 3 take a close look to the defect and normal samples. For classification, a tiny deformation should be detected as shown in the first image in Fig. 3.

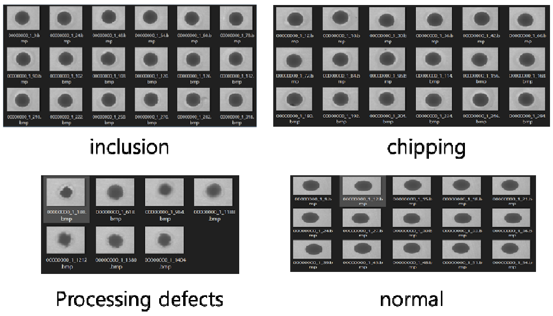


Fig. 2. The examples of various normal and defect classes

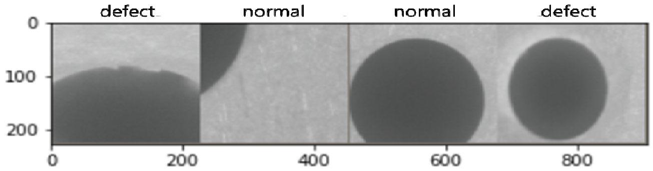


Fig. 3. The Cropped Ceramic Data Samples.

3 Defect Classification

We use pretrained Resnet18[3] model trained by ImageNet[4] datasets with 1000 classes. Our data does not have rich texture information and the number of samples are limited. Therefore, we adopt the pretrained model. Moreover, our problem is classification and ImageNet is close to our problem space.

By adopting transfer learning, we remove the last fully connected layer with 1000 output channels and add fully connected layer with two output channels for binary classification: normal and defect. Then, we trained only the last layer using cross entropy criterion. We use Adam optimizer with learning rate = 0.001.

4 Experimental Results

For biniary classification, normal data is 463 (232 for training, 231 for test) in total and defect data is 726 images (372 for training and 354 for test). In Fig. 4. the training loss continuously went down. The training accuracy kept rising up to 90.73%. The test accuracy is 98.46% which is higher than training accuracy. This meant our model is not overfitting. Fig. 5. Showed the qualitative results.

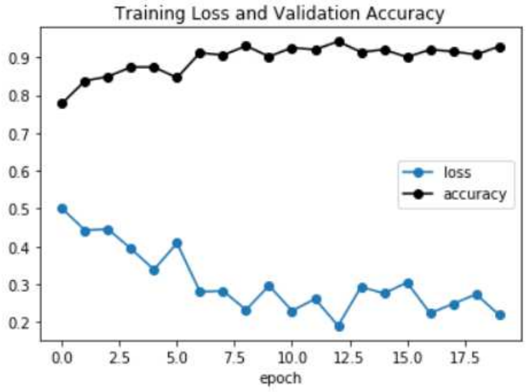


Fig. 4. The train loss and accuracy graph.

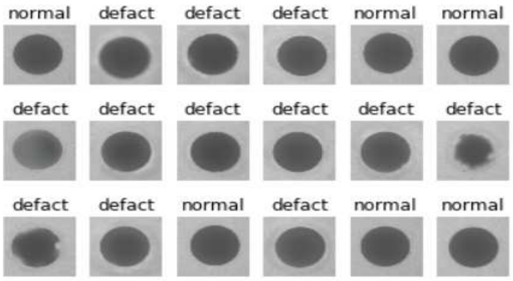


Fig. 5. The qualitative test results

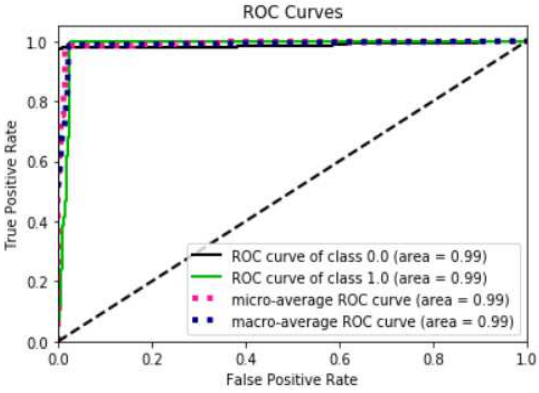


Fig. 6. Receiver Operator Characteristic (ROC) curve

Table 1. Ceramic defect classification Results

	precision	Recall	f1-score	support
Class 0(defect)	1.00	0.97	0.99	354
Class 1(normal)	0.96	1.00	0.98	231

5 Discussion

We presented defect classification results by using ResNet18 pretrained model, which is a well-known deep learning algorithm. Because of ill-conditioned data statistics, we did binary classification and achieved 98.46% classification accuracy. For the future work, we classify the defects in details and detect the defect regions. Moreover, we generate insufficient data such as crack by employing GAN techniques. In addition, we try to develop a solution to data imbalance.

Acknowledgement

This material is based upon work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under i-Ceramic platform construction project - i-Ceramic manufacturing innovation platform. No. 20004367, 'Development of Cloud Big Data Platform for the Innovative Manufacturing in Ceramic Industry'

References

1. Bashkansky M., Duncan, M.D., Kahn, M., Lewis D. III, Reintjes J.F.,: Subsurface defect detection in ceramic materials using low coherence optical scatter reflectometer, Proceeding of a Joint Conference, Mobile, Alabama (1996)
2. Ahamad, N. S., Rao, J. B.: Analysis and Detection of Surface Defects in Ceramic Tile Using Image Processing Techniques, Microelectronics, Electromagnetics and Telecommunications, Lecture Notes in Electrical Engineering 372 (2016)
3. He, K., Zhang, X., Ren, S. Sun, J.: Deep Residual Learning for Image Recognition, Proceeding of CVPR (2015)
4. Deng, J. and Dong, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L:ImageNet: A Large-Scale Hierarchical Image Database, Proceeding of CVPR (2009)

Process Quality Estimation using Statistical Process Features

Pheng Tola, Ga-Ae Ryu, Kwan-Hee Yoo*

Dept. of Computer Science, Chungbuk National University, South Korea
tolapheng99@gmail.com, gary, kyoo{@chungbuk.ac.kr}
*Corresponding Author

Abstract. Every manufacturing industry struggles to improve product quality during the production process. Higher quality production is considerably for generating satisfactory profits. In this paper, we propose a composite statistical process analysis such as statistical process control, process capability analysis, and process trajectory outlier detection, which is used to estimate abnormal production processes and to improve the quality of products in the manufacturing industry. The proposed method is exemplified by a real case study, smart manufacturing system to verify the quality improvement. As a result, product quality has significantly increased.

Keywords: Process Quality, PQE, statistical process features

1. Introduction

In recent years, manufacturing processes are getting more and more complex. The high complexity of manufacturing industry and the continuously growing large amount of data drive to excessive demands on the manufactures for process monitoring and data analysis. Improving product quality has become the most important part of the manufacturing industry. A manufacturing producer can dominate the other competitors when the customers can receive high-quality products. Recent decade, many studies have proposed their methods to improve production process quality control and to increase the speed of producing products with the best quality. The delivery of a good final product is an important part of the manufacturing process [1]. It seems like a difficult task with lots of challenges of balancing in manufacturing processes [2]. This paper is organized in the following sections after this introduction. Section two discusses the related study. Section three describes the proposed method of estimation of process performance. Section four shows the experimental result. Finally, the paper concludes with the results of this study.

2. Related Study

There are copious studies of statistical process control (SPC) and process capability analysis (PCA) to determine whether the process is in control that can produce high-quality products. In [3], Zheng used Bayesian statistics to analyze the historical data for quality control and measurement. They proposed the Bayesian statistic to apply in the equipment maintenance process in manufacturing quality control. Their research showed that the capability of control chart based on the Bayesian Statistic method is good to maintain quality control. Burlikowska, et. al [4] proposed control charts such as X and R charts, and process capability to estimate the production process in the polish industry to avoid the defective product in the production lines. Their result described that the control charts type X and R charts are not capable, and the production process is needed to be improved by the manufacturers to prevent product failure.

3. Proposed Methods

The adopted methodology in this paper is Process Quality Estimation (PQE) which is the combination of SPC X & R charts, PCA Cp & Cpk and process trajectory outlier detection (PTOD). The goal of PQE is to monitor and estimate the process quality in order to prevent the defective product. Many previous results have been proposed to monitor and measure production process quality such as SPC X & R charts and PAC Cp & Cpk. Both methods are related that help to understand the production process. The PCA demonstrates how capable of the production process is producing the product, and it checks whether the process is in control with specification requirements designed by the manufacturers before starting to produce a product in the machine [4-5]. With this study, we use Cp & Cpk to indicate whether the production process is meeting the specification limits. Process capability (Cp) is calculated using the specification limit and the standard deviation [10]. Process capability index (Cpk) is calculated using specification limits, the standard deviation, and process mean. SPC X & R charts are used separately to measure production process quality. We need to calculate three straight lines to measure the production process such as Upper Control Limit line (UCL), Lower Control Limit line (LCL), and CL (Central Line). X chart is the mean value of a production process or range value, and R chart is the standard deviation of the production process [6]. However, we combined one more method PTOD to improve the estimation of a production process with high accuracy and decrease the number of defective products. There are several processes to produce one product of a machine. One process is defined as one process trajectory. Trajectory abnormal detection (TAD) is so valuable analysis method in different scenarios, trajectory abnormal detection has been considered such as abnormal trajectory, abnormal sub trajectory, abnormal road segment, and abnormal event, abnormal moving object, etc. We applied distance-based to detect the outlier in process trajectories, so if any points of a process trajectory are an outside threshold, it is considered an outlier. Then We computed score from each point of X & R values, Cp & Cpk measurement results, and PTOD scores. Then we estimate the PPQ by calculating with mathematical equations which are described in subsection 3.1.

3.1 Process Quality Estimation

PQE is a methodology to estimate the production process quality for smart manufacturing. It is comprised of PTOD, SPC X & R charts, and PAC Cp & Cpk analysis result. The score of PQE is computed every 30 minutes time interval. After computing score from three factors, we estimate the PQE by the following formula:

$$PQE = 1/n(\sum_{i=1}^{n} T_i + R_i + X_i \ + Cp_i \ + Cpk_i)$$

Where n is the number of products, T_i is the score of PTOD. If the PTOD has no process points outside the threshold, so the score is one, otherwise zero. Moreover, R_i is the value of the R chart, and X_i is the value of the X-bar chart. Similarly, the scores for the X chart and the R chart are one, if the data points are inside UCL and LCL. In the case, the scores of Cp_i and Cpk_i are smaller than one, then the process is incapable [7]. However, if the calculation result is greater than two, the scores of Cp_i and Cpk_i are two. If the result is smaller than two and greater than zero, the scores of calculating results will be assigned to the score of Cp_i and Cpk_i.

4. Experiment Results

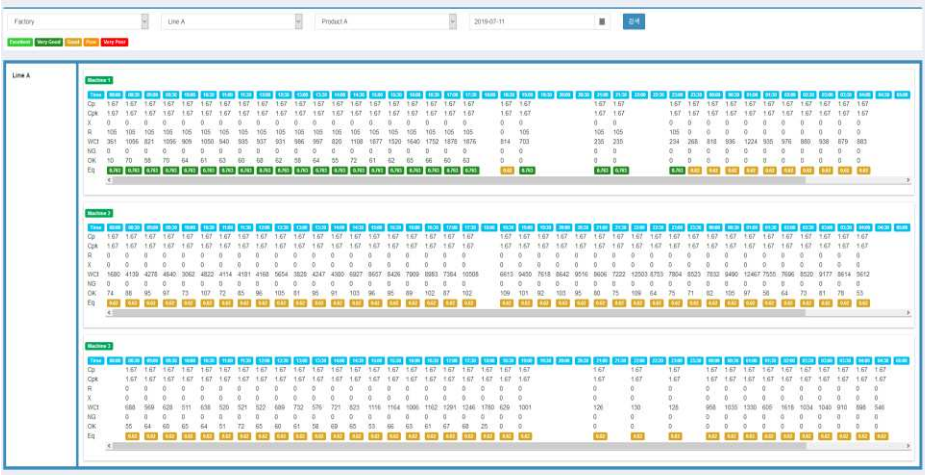


Fig. 1 Process Quality Estimation Results

In this section, we describe the experimental results of PQE. In Figure 1 shows the process quality estimation, the data is collected from running three machines of one line from 8 am until the next day 8 am. We analyze and estimate process quality every thirty minutes. The results of the estimation are classified into five classes. If the PQE is equal 1.00, we estimate the process is an excellent performance. If the PQE is smaller than 1.00 and greater or equal than 0.75, we estimate the process is a very good performance. If the PQE is smaller than 0.75 and greater or equal than 0.50, we consider that the process is a good performance. Whereas if the PQE is smaller than 0.50, we consider that the process is poor or very poor. This is the dangerous results that the manufacturers need to check the machines to find the problem and solve it before defective products occur. As the results, all machines were performing well, and almost of products met the design specification without any defective products.

5. Conclusion

In this paper, we proposed a composite analysis method which consists of PTOD, SPC X & R charts, and PCA Cp & Cpk to estimate the process quality which could improve product quality and prevent product before failure occur. With the three comprised methods, we compute a score from each analysis. Then we use the computation score to calculate PQE to estimate the quality of processes. As a result, the number of defective products has considerably decreased.

Acknowledgments. This material is based upon work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program, No.10082578, 'Development of intelligent operation system based on big data for production process efficiency and quality optimization in non-ferrous metal industry.

References

1. T. A. Ryabchik, E. E. Smirnova, and M. l. Lukashova, “Manufacturing Processes Quality Control as a Main Factor of Performance Enhancement in Industrial Management” Saint Petersburg and Moscow, Russia, Russia, Jan 2019.
2. L. Ying, and W. Sujie, “The construction study of quality cost evaluating system based on advanced manufacturing environment” Beijing, China, Oct 2009.
3. Y. Zheng, Q. Geng, and R. He, “The Application of Control Chart Based on Bayesian Statistic in Equipment Maintenace Quality Control” Chengdu, China, July 2013.
4. M. D. Burlikowska, “Quality estimation of process with usage control charts type X-R and quality capability of process Cp, Cpk” Vol.162-163, pp. 736-743, May 2005
5. M. Yadav, M. K. Sain, and D. Joshi “Analysing the Process Capability of Carburettor Manufacturing” Vol.8, 2018
6. C. E. Okorie, O. Adubisi, and O. J. Ben “Statistical Quality Control of the Production Materials in Line Leager Beer” Vol.2, pp. 69-73, April 2017
7. M. Spinola, M. Pessoa, A. Tonini, “The Cp and Cpk Indexes in Software Development Resource Relocation”, Portland International Conference on Management of Engineering & Technology, October 15, 2017

An Emotion Graph Generation Scheme for Actor in Movie Using Face Expression Recognitions

Sungho Han, Kyoungsik Hong, Sang-Won Lee, Kwangho Song, Yoo-Sung Kim

Department of Information and Communication Engineering
Inha University
Incheon 22212, Korea
yskim@inha.ac.kr

Abstract. In this paper, we propose an emotion graph generation scheme for actors in movie using face expression recognitions, to help analyze the story structure of the input movie. For this purpose, the specified actor is searched in the frames of the movie, and his/her emotions are analyzed by the face expression recognitions, and finally generated and displayed the emotion graph. Also, we develop a prototype system of the proposed emotion graph generation scheme. From the experiments with sample movies, we can see that the emotion graphs generated by the prototype system are useful to represent the emotion changes of actors and to understand the movie stories well.

Keywords: Emotion graph, movie actor, face detection and alignment, face matching, face expression recognition.

1 Introduction

Analyzing emotions of actors in a movie is very helpful to understand the story of the movie, since the emotions of movie actors varies according to the movie story [1-3]. Since the previous study([3]) which proposes a story analysis method of the novel by analyzing the occurrences of words in the novel text, we can get the story information of the movie when the script of movie is in hand to input the scheme. However, in general, since we cannot get the scripts of movies easily, we cannot use the method to analyze the story of movie. And up to our knowledges, there is no system or service by which the story of movie is analyzed from the movie film itself. On the other hand, there have been many studies on analyzing face expressions from video or images[4-7]. So, in this study, we propose a novel system which can analyze the emotions of movie actor by analyzing his or her face expressions and can generate finally the emotion graph for movie actors from the movie film.

The proposed emotion graph generation scheme to support understanding the movie story accepts and processes the movie film with the face image of the specific movie actor whose emotions are analyzed from the movie in the following procedure. First, from the input movie film, the frames in which the face of the specific movie actor is is detected. The face area detected is cropped and transformed into the corresponding face image looking ahead to enhance the emotion recognition accuracy.

And the emotions of the specific actor varying according to the story in the movie are recognized and represented as the emotion change graph. In other words, first sample frames from the input movie film are appropriately selected and human faces are detected and located in the sample frame. And then to correctly identify only the face of the required emotion analysis actor and to correctly analyze the emotion of that actor, the detected face is transformed into the corresponding face looking ahead. Otherwise, the other frames in which no face matched to the required actor is detected or the matched face can not be transformed are ignored in the next processing. From the transformed face image, emotion is analyzed by using a convolutional neural network architecture into 8 categories, happy, sad, disgust, contempt, surprise, anger, fear, neutral with the associated probabilities. Among the 8 emotions, the one with the highest probability is determined and the emotion graph is generated from these analyzed face expressions. By using the generated emotion graph for the required actor in movie, users can understand easily the story of the given movie.

The rest of this paper consists as follows. Section 2 briefly introduces the previous works on emotion analysis from face images. In section 3, as the main components of the emotion graph generation scheme, face detection, face alignment, similarity computation for face matching, and emotion classification and emotion graph generation are described. Section 4 also introduces the graphical user interface of the developed prototype system and the emotion analysis result by the prototype system is discussed. Finally, section 5 summarizes this study as the conclusion.

2 Related Works

Many methods for recognizing emotions from facial images have been studied. These studies can be divided into model-based methods, image-based methods, and combining these two methods[4]. The model-based recognition method recognizes emotion based on extracted facial features or outlines[4,5]. The image-based method recognizes emotion using the overall brightness of the face image and the features of the eyes, nose and mouth area [4,6]. Combining these two methods extract facial features in different ways and then recognize facial emotions using a neural network or Hidden Markov Model (HMM).

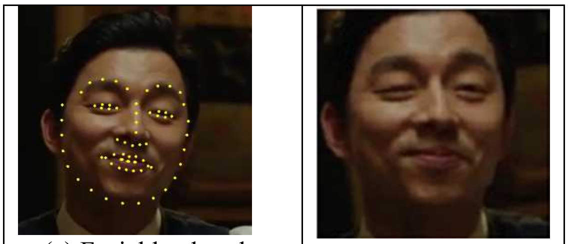
Recently, [4] proposed an image-based method that extracts features by applying LBP(Local Binary Pattern) to face images and classifies them into 4 types of emotions, normal, smile, surprise, and angry using SVM(Support Vector Machine). Also recently, instead of extracting and using the specific features from the face image using image processing, the methods that input face image directly to the convolutional neural network to find the optimal convolution layer and fully connected layer structure for emotion analysis have been proposed[7].

However, these previous studies only analyze emotions from frontal face images taken in limited environments. Therefore, it is difficult to recognize a specific actor and analyze emotions based only on the actor's face image. Because, there are various poses and various face angles that actors naturally play in movie. In this study, we propose a system that extracts the actor's face from image frames that specific actor appear and classifies emotions.

3 Emotion Graph from Face Expression Recognitions

This section describes the process of finding the frame images in which the required actor appears, analyzing emotions and generating an emotion graph according to the movie story. First, in order to use only movie frames with human faces as analysis targets, human faces were recognized using Haar Cascade and only frames where faces were detected were sampled. Haar like feature is related to the contrast of brightness and darkness of specific object. So for human face, the common observation that the region of the eyes is darker than the region of the nose can be used. That is, there are shadows around the eyes and the nose receives light. Therefore, the rectangles with these characteristics respectively are created and used to find the human face area that matches to the rectangles [8]. Sometimes, since face detection only using the Haar Cascade method produces errors, to prevent this problem, another confirmation step using the HOG is used in this study. The HOG is a pattern descriptor that describes the degree and direction of pixel value changing from one point to another in the image[9]. Using this method, we can accurately find faces in images by comparing HOG features that often appear on faces with HOG features extracted from current images.

The next phase is for face alignment. The actor's face found in the above process may not be a frontal view of the face. This may reduce accuracy in determining whether the detected face is of the required actor and in analyzing the emotion from the facial expressions. Therefore, in this study, we use the affine transformation that converts the detected face into a frontal view of the face as the second phase for emotion recognition. To perform the affine transformation, facial landmarks like eyes, nose, mouth and face circumference are detected first in the face image. We used DLIB to find these facial landmarks and DLIB uses a regression tree for training [10]. The Figure 1-(a) is the detected facial landmarks, and Figure 1-(b) is a frontal view of the target actor's face converted from the detected face based on the facial landmarks.



(a) Facial landmarks

(b) Aligned face

Figure 1. Transform face into a frontal view of the face

The third phase is to check whether the recognized faces are of the required actor for the emotion analysis. The VGG Face model was used to determine the similarity between the inputted face image of the specific actor for emotion analysis and the face found in the sampling frame of the movie[11]. The VGG Face model embeds face images into vectors using weights learned through the Triplet Loss function. Based on this model, vectors are generated from the face image of the target actor and the detected face image in the movie, respectively. Then, the similarity between the vectors was evaluated by Cosine similarity. If the similarity is above the threshold, they are considered as the same person.

Finally, emotional graph is generated by analyzing emotions only for the faces of the target actor in the movie. Emotion is classified into one among 8 categories such as happy, surprise, sad, anger, disgust, fear, contempt, and neutral. For classification of emotions, classification model was trained through the Fer+ dataset[12]. The model used for learning is a CNN-based model with the structure shown in Figure 2 below[13].

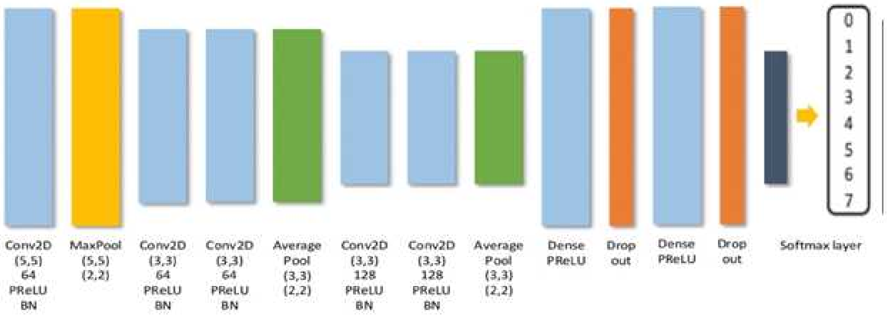
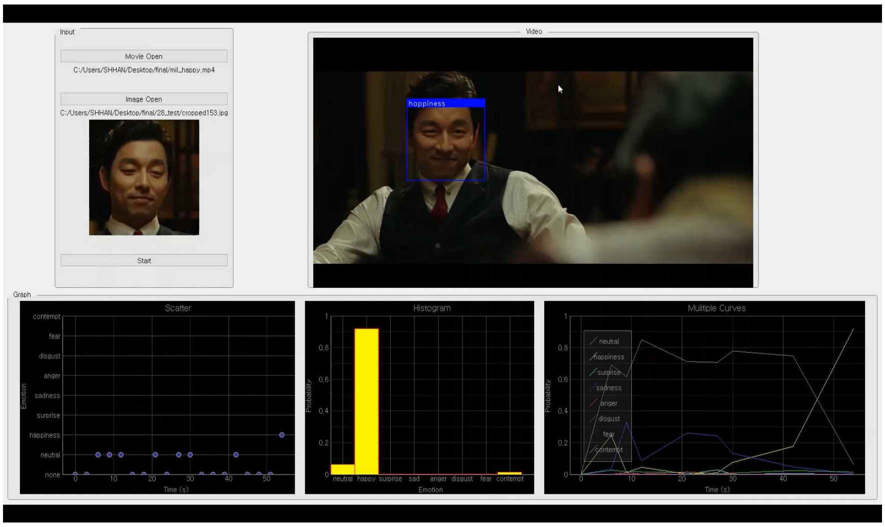


Figure 2. CNN-based network architecture for emotion classifier

4 Prototype Development of Emotion Graph Generation System

We developed a prototype system for visualizing and analyzing emotion changes of the specific actor in a movie. Figure 3 shows the main GUI(graphical user interface) of the developed prototype system. First, as shown in the upper left of the main GUI, the movie film file to analyze and the face image of the target actor are input and specified. Then, using the sampled frame and the face image of the target actor, the analyzed emotion are graphed in three ways at the bottom of the GUI. In addition, the upper right of the GUI shows a sampled frame image to check real-time progressing. In this window, the face of the target actor is marked with a blue box and the emotion of the face is displayed on it. And if the detected faces are of other actors not the specific actor, they are marked with yellow boxes to distinguish from the face of the specific actor.

As the emotion change graph for the specific actor, there are three emotion analysis graphs as shown at the bottom of Figure 3. The first graph plots the emotion that have a high probability as points over time. The second graph shows the probability of each emotion as a histogram. The third graph shows the probability of eight emotions as a multi curve graph over time.



C:/Users/SHHAN/Desktop/final/mil_happy.mp4

C:/Users/SHHAN/Desktop/final/28_test/cropped153.jpg

Figure 3. GUI of emotion analysis graph generation system

As example of an actor's emotional analysis graph, Figure 4 shows the result of showing the emotions over time as actress Yeom Jung-ah in the Korean movie of '완벽한 타인'. One frame every three seconds was extracted as sample frames and analyzed. Each emotion is marked in the different colors.

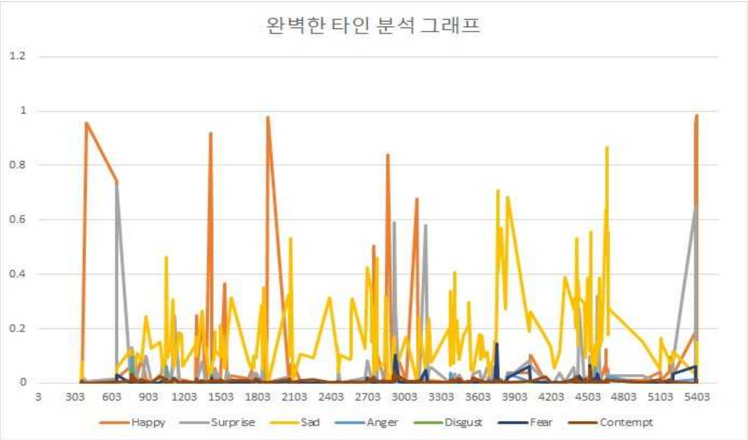


Figure 4. Example of an actor's emotion change graph

5 Conclusions

In this study, the method for generating emotion graph from face expression analysis for the required actor in movie is proposed and a prototype system is developed. In the proposed method, the emotion of the specific actor is analyzed from the face expression recognition in 4 steps as follows, human face detection and localization step in the sampled frames from the input movie film, face alignment step to transform the detected face image into the corresponding face looking ahead, identify and confirm the detected face image to the face of the required actor, and finally generating emotion graph by representing the emotions of the required actor in the movie according to the movie play time. From the analysis of the main actor in the sample movie, the developed prototype system with the proposed emotion graph generation scheme is known to be very helpful to understand the movie story.

References

1. Kim, N, Jung, G., Lee, M., Effects of an Actor and an Actress on Viewers' Intentions to Watch Romantic Comedies, Journal of Marketing Management Research, Vol. 16, No. 1, pp. 125-143, (2011)
2. Chung, J., Eoh, J., A Marketability Forecasting Model for Story-based Content, Journal of Korean Marketing Association, Vol. 25, No. 2, pp. 65-88, (2010).
3. Reagan, Andrew J., et al. The emotional arcs of stories are dominated by six basic shapes, EPJ Data Science, Vol. 5, No. 1, pp. 1-12, (2016).
4. Won, C., Recognition of Facial Emotion Using Multi-scale LBP, Journal of Korea Multimedia Society, Vol. 17, No. 12, pp. 1383-1392, (2014).
5. Gao, Y., Leung, K.H., Hui, S., Tananda, Facial Expression Recognition from Line-based Caricatures, IEEE Trans. On Systems, Man, and Cybernetics-Part A: Systems and Humans, Vol. 33, No. 3, pp. 405-412, (2003).
6. Kim, H., Joo, Y., Park, J. Lee, J., Cho, Y., Development of Emotion Recognition System Using Facial Image, Journal of Korean Institute of Intelligent Systems, Vol. 15, No. 2, pp. 191-196, (2005)
7. Pramerdorfer, C., Kampel, M., Facial Expression Recognition using Convolutional Neural Networks: State of the Art, arXiv preprint arXiv:1612.02903.
8. Viola, P., Jones, M., Rapid Object Detection using a Boosted Cascade of Simple Features, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, (2001).
9. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D., Object Detection with Discriminatively Trained Part-Based Models, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627-1645, (2010).
10. Kazemi, V., Sullivan, J., One Millisecond Face Alignment with an Ensemble of Regression Trees, 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, pp. 1867-1874, (2014)
11. Taigman, Y., Yang, M., Ranzato, M., Wolf, L., DeepFace: Closing the Gap to Human-Level Performance in Face Verification, 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, pp. 1701-1708, (2014).
12. Emad, B., et al., Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution, Proceedings of the 18th ACM International Conference on Multimodal Interaction, (2016).
13. https://github.com/Wonjuseo/Facial_Expression.

A Graphics Pipeline Architecture with OpenCL

Mingyu Kim[1], Nakhoon Baek[1, 2, 3]

[1] School of Computer Science and Engineering, Kyungpook National University,
Daegu 41566, Republic of Korea
[2] Software Technology Research Center, Kyungpook National University,
Daegu 41566, Republic of Korea
[3] dassomey.com Inc., Daegu 41566, Republic of Korea
oceancru@gmail.com

Abstract. Software graphics pipeline like CUDARaster has high flexibility and performance. However, CUDARaster has some limitations including that it only works on CUDA and executing it on recent architectures fails. In this paper, we implement software graphics pipeline architecture based on OpenCL, aimed at various platform support. We evaluate performance in two different environments. Then, we compare performance with CUDARaster and our work in various platforms. Our pipeline has a rendering result better than CUDARaster, but there was some lack of optimization.

Keywords: parallel computing, graphics pipeline, OpenCL

1. Introduction

Nowadays, GPGPU technology is growing rapidly with the emergence of artificial intelligence and big data. OpenCL [1] is a new industry standard for task-parallel and data-parallel heterogeneous computing on a variety of modern CPUs, GPUs, DSPs, and other microprocessor designs. CUDA [2] is a general-purpose parallel computing platform and programming model that leverages the parallel compute engine in NVIDIA GPUs to solve many complex computational problems in a more efficient way than on a CPU. OpenCL and CUDA allow the GPU to be operated in compute mode.

By comparison, graphics pipeline is evolving slowly and still restrictively allow programmable stages. Although built-in hardware implementations can achieve powerful performance and efficiency, programmer cannot handle intermediate process such as rasterization. CUDARaster [3] is the highest performance software graphics pipeline. However, since it is implemented in CUDA, it is only for NVIDIA graphics cards. It also relies on optimizations targeting the outdated Fermi architecture, execution on architectures more recent than Kepler fails [4]. Therefore, we implement graphics pipeline architecture with OpenCL so that it can be run on various platforms from the latest graphics card to integrated graphics card. We support more functions than rasterizer and guarantee the input order.

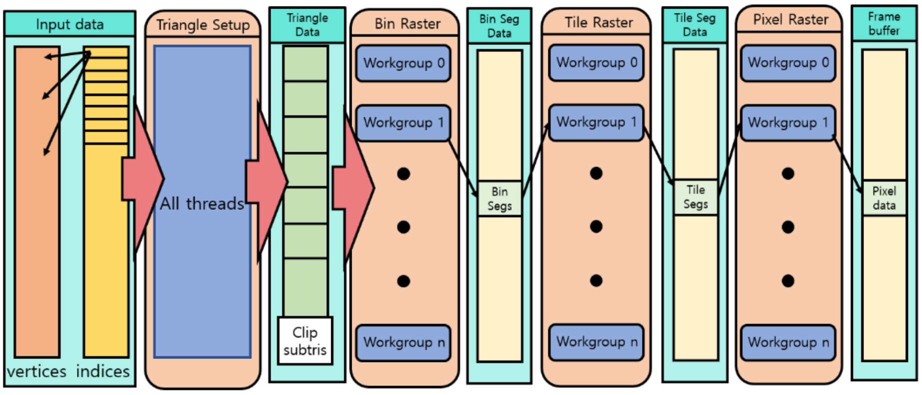


Fig. 1. The overall execution sequence of rasterizer implemented in this paper. Each stage is discussed in detail in Section 2.

2. Implementation

2.1 Implementation Strategy

We followed chunker-style pipeline structure of CUDARaster. Overall execution sequence is shown in Fig.1. We use OpenCL-OpenGL interoperation function to draw result of rasterizer. OpenCL-OpenGL interoperation is supported above OpenCL 1.2 version. So, we targeted OpenCL version 1.2.

The biggest different between CUDA and OpenCL is that CUDA warp, which determines logical threads that run on same concurrency, is a set of 32 threads based on NVIDIA graphics card. Whereas, OpenCL wavefront is a set of 64 threads based on AMD graphics card. Therefore, we defined threads multiple of 64 when defining number of work-items for local workgroup. Also, maximum number of work-items in a global and local workgroup is different in graphics card architecture, we should consider that either.

Since CUDARaster has limited platform to NVIDIA GPU for optimization, it has implemented using special Warp Vote Function such as __any, __all, and __ballot supported by NVIDIA, and a low-level instruction set architecture. However, in this implementation, we aimed to support various platform, so we focused on porting CUDARaster CPU method to GPU method as much as possible. Optimization refers to CUDARaster GPU method.

2.2. Triangle Setup

In this step, each thread deals with one triangle. We divide local workgroup into 64 work items. The number of global work item is in proportional to total number of triangles we have to rasterize.

First, each thread gets input vertex position values along the vertex index of the triangle. Second, we perform view frustum culling. Then, we fill triangle header and data if triangle is inside depth range and visible. Finally, if triangle needs to be clipped to view frustum, we allocate space and fill sub-triangle header and data. We use atomic functions when allocate space to set aside race conditions.

2.3 Bin Raster

Bin Raster uses triangle data from Triangle Setup to determine which triangle belongs to the bin. Each local workgroup works on one bin. And since maximum number of work-items in a workgroup is 256, one local workgroup has 256 threads. Y and Z coordinates of the global workgroup are mapped to x index and y index of bin. One bin has 128x128 pixels.

Each thread composes local array after assembling triangle data. Then, perform intersection check of the triangle and the bin. In this stage, we perform odd-even sorting to ensure the order of the input data and repeat until all the triangle data are read.

2.4 Tile Raster

Tile Raster uses segment data from Bin Raster to determine which segment belongs to the tile. Each local workgroup works on one tile. The x and y coordinates of the global workgroup are mapped to x and y index of tile. Each tile has 8x8 pixels, because NVIDIA Geforce GTX 960 limits maximum global workgroup size to 1024x1024x64.

Each thread composes local array after assemble bin segment data. Then, perform intersection check of the triangle segment and the tile. Like previous stage, perform odd-even sorting to ensure the order of the input data and repeat until all the bin segments are read.

2.5 Pixel Raster

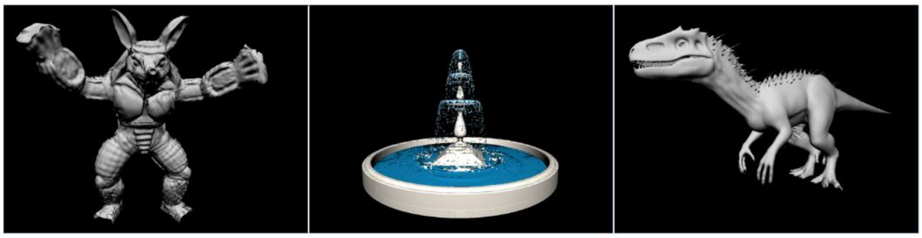
Pixel Raster uses segment data from tile raster stage to determine with segment belongs to the pixel. Each local work group works on one tile and each thread works on one pixel. The x and y coordinates of the global workgroup are mapped to x index and y index of tile. This applies for the same reason of the tile raster stage.

In OpenCL, it does not support specific texture format like GL_RGB. Also, each texture has various size. We address this problem to link all texture data as one dimension of char type array. And next stores the texture information (type, width, height) in the constant memory. We also store the light information (position, ambient, specular, diffuse) and material information (ambient, diffuse, specular, shininess) in the constant memory.

3. Performance Evaluation

We evaluated performance in two platforms. First PC environment is Intel Core i5-3550 CPU and 8GB of RAM. The operating system is Windows 7. Set up graphic card is Geforce GTX 660, GTX 960. Second PC environment is Intel Core i5-6500 CPU with 16GB of RAM. The operating system is Windows 10. Attached Graphics card is Radeon RX 570.

We evaluate performance of three scenes to compare performance per the number of faces. The first scene Armadillo has 212,574 faces without texture. Second scene Fountain has 187,410 faces and textures. Last scene Dinosaur has 121,858 faces without texture. Note that we have to make sure that the texture is applied properly. All Rasterizer should apply depth test and phong shading but MSAA. Screen Size is 1024x768. We evaluate performance per each stage except vertex shader stage in our rasterizer and CUDARaster.



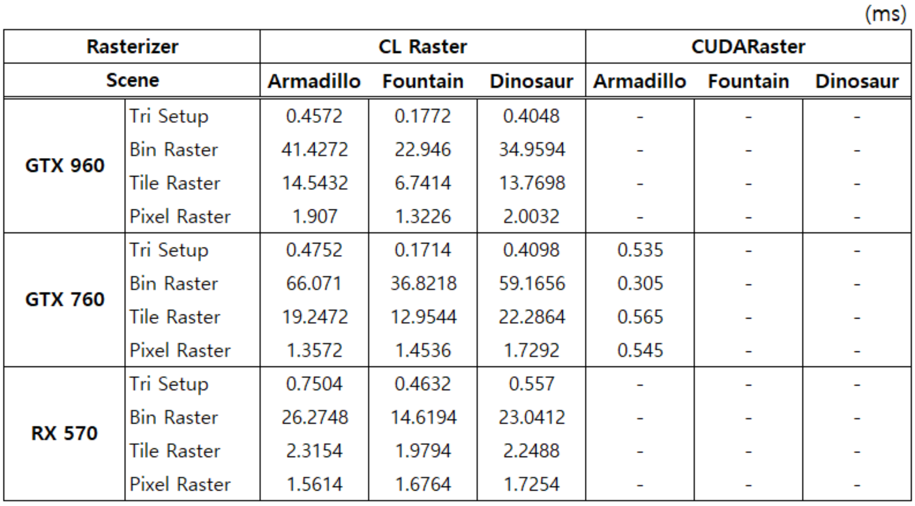
Armadillo
212,774 tris

Fountain
187,410 tris

Dinosaur
121,858 tris

Fig. 2. Rendering results of our works. Phong shading and depth test applied.

Table 1. Performance comparison between CUDARaster and our software pipeline with various platform. We evaluate performance except vertex shader and buffer swap.



(ms)

Rasterizer		CL Raster			CUDARaster		
Scene		Armadillo	Fountain	Dinosaur	Armadillo	Fountain	Dinosaur
GTX 960	Tri Setup	0.4572	0.1772	0.4048	-	-	-
	Bin Raster	41.4272	22.946	34.9594	-	-	-
	Tile Raster	14.5432	6.7414	13.7698	-	-	-
	Pixel Raster	1.907	1.3226	2.0032	-	-	-
GTX 760	Tri Setup	0.4752	0.1714	0.4098	0.535	-	-
	Bin Raster	66.071	36.8218	59.1656	0.305	-	-
	Tile Raster	19.2472	12.9544	22.2864	0.565	-	-
	Pixel Raster	1.3572	1.4536	1.7292	0.545	-	-
RX 570	Tri Setup	0.7504	0.4632	0.557	-	-	-
	Bin Raster	26.2748	14.6194	23.0412	-	-	-
	Tile Raster	2.3154	1.9794	2.2488	-	-	-
	Pixel Raster	1.5614	1.6764	1.7254	-	-	-

4. Results and Discussion

Because CUDARaster only works on GTX 760 with scene 1, we had to compare only the results of Scene 1. Triangle Setup stage is faster than CUDARaster. Pixel Raster stage is slower than CUDARaster but there is no large variation with regards to the number of faces. Also, as we can see in the Table 1, as the number of faces increase more, Bin Raster and Tile Raster spend much more time.

When we ran the Tile Raster on the RX 570, it is about six times faster than other graphics cards. However, we found that the Triangle Setup stage on the RX 570 has lower performance than other graphics cards.

Though overall performance is not better than CUDARaster. Because this time, we took more focus on getting the right rendering results and provide platform-free software rasterizer.

Currently, there are few devices that support OpenCL version 2.0 or higher. However, if many devices support OpenCL version 2.0 later, we can expect performance improvement in atomic operations and to use additional features. In addition, if OpenCL support a low-level instruction set like CUDA, we can optimize our work better. We can apply multi-sampling feature in the future to get better rendering quality. We will also get better results later by optimizing the task distribution or parallel algorithm of our work.

5. Conclusion

So far, Software Graphics Pipeline has been developed and studied mainly in CUDA. CUDARaster has disadvantages of not being supported on various platforms such as Radeon graphics card and Intel graphics card. In this study, we showed software graphics pipeline which supports various platforms and supports more functions than rasterizer. With the same approach, we expect wider range of research on various platforms in the future for the software graphics pipeline with OpenCL.

In this time, we use OpenGL-OpenCL interoperation function to render our rasterizer result. But if OpenCL supports some GL functions in own library, we will be able to create software graphics pipeline with only OpenCL.

Our work has performance problem and supports only fixed function pipeline. So in future, we plan to optimize our work and support various pipeline extensions. We can also modify the existing pipeline to find ways to render more efficiently in software graphics pipeline.

Acknowledge

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant 2016R1D1A3B03935488 and 2019R1I1A3A01061310).

Reference

1. Stone, J., Gohara, D., Shi, G.: OpenCL: A Parallel Programming Standard for Heterogeneous Computing Systems. In: Computing in Science & Engineering, vol. 12, pp. 66--73. IEEE Press (2010)
2. NVIDIA.: CUDA C Programming Guide. NVIDIA Corporation (2012)
3. Laine, S., Karras, T.: High-Performance Software Rasterization on GPUs. In: Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics, pp. 79--88. ACM, New York (2011)
4. Kenzel, M., Kerbl, B., Schmalstieg, D.: A High-Performance Software Graphics Pipeline Architecture for the GPU. In: ACM Transactions on Graphics, vol. 37, Issue, 4. ACM, New York (2018)

Full-color Holographic System featuring Compressed Point Cloud Gridding for representation of real 3D objects

Yu Zhao[1,2], Ki-Chul Kwon[2], Yan-Ling Piao[2], Munkh-Uchral Erdenebat[2], Kwan-Hee Yoo[3*], and Nam Kim[2*],

[1] School of Information Engineering, Yangzhou University, Yangzhou 225127, China
[2] School of Information and Communication Engineering, Chungbuk National University, Chungbuk 28644, South Korea [3] Dept. of Computer Science, Chungbuk National University, Chungbuk 28644, South Korea
namkim@chungbuk.ac.kr

Abstract. A multiple-camera holographic system featuring non-uniform sampled 2D images and compressed point cloud gridding is developed. High quality digital single lens reflex (DSLR) cameras are used to acquire depth and color information from real scenes, and then virtually reconstruct uniform point cloud by using a non-uniform sampled method. Compressed point cloud gridding (C-PCG) method is proposed to accelerate the calculation speed in GPU for the full-color holographic system. The feasibility of our method was confirmed both numerically and optically.

Keywords: Holography, Computer holography, Color holography, Holographic display

1 Introduction

Real object-based methods have become more attractive over the past few years, because they are simple to apply and afford efficient natural visualization of such objects [1]. A recent algorithm applied the point cloud gridding (PCG) method to accelerate CGH generation from real objects [2]. Depth camera was employed to simultaneously acquire depth and color data from real scenes and a color point-cloud model then extracted. The low resolution of the depth camera limits the quality of the reconstructed image. Therefore, a full-color holographic system featuring non-uniform sampled 2D images and compressed point cloud gridding is developed for representation of real 3D objects.

2 Proposed system

There are five main steps to the proposed full-color hologram generation method: acquisition, point cloud generation, compressed point cloud gridding(C-PCG) process, hologram-generation, and reconstruction, as shown in Fig. 1. The procedure of full-color holographic system with a real 3D object can be described as follows:

(1) Multi DSLR cameras are scanned horizontally as shown in Fig. 1. Acquire 33 × 3 viewpoints light field 2D images from DSLR cameras.
(2) Acquire uniform point cloud from 2D images, then identify the ROI process.

(3) Compressed calculation is adopted for the z coordinates to reduce the number of depth layers. And Ensure that the hologram has the desired resolution by stretching the coordinates of the point cloud, and then relocate the sub-layers of the 3D object to eliminate repeated values.

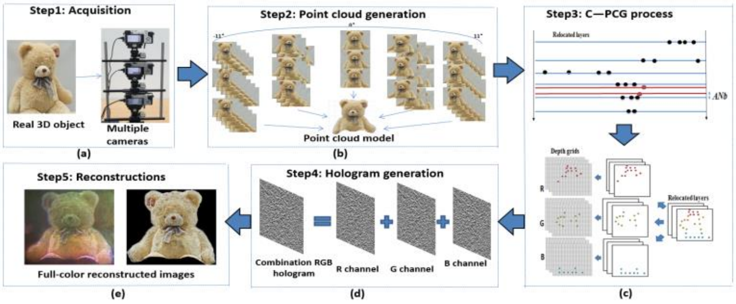


Fig. 1. Outline of the proposed full-color holographic system

We employed numerical simulations and optical experiments to evaluate the performance of the proposed method. The experiment was implemented in MATLAB 2017b and run on a Windows 10 64-bit PC with 8 GB RAM, and an NVIDIA GTX 660 GPU. DSLR cameras (Sony α6000, resolution 6000 × 4000) are used to capture the real objects. Numerical simulation and optically reconstructed images of the proposed full-color holographic system are shown in Fig. 2. The reconstructed images have higher quality than those generated by conventional methods. When the GPU is used, the hologram generation speed of proposed method is enhanced 3.9~5.5 fold in comparison with wave-front recording method.

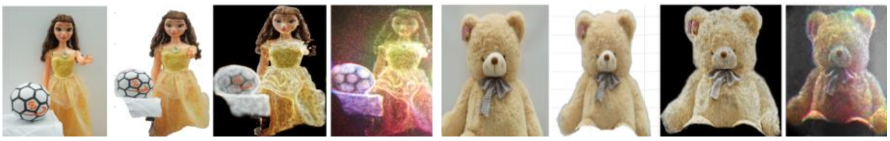


Fig. 2. Numerical and optically reconstructed images from real objects.

3 Conclusion

In this paper, a full-color holographic system is developed for the representation of real 3D objects. The depth and color data of the real scene were simultaneously acquired through the multiple DSLR cameras. Real 3D object can be easily encoded into CGHs with the proposed algorithm. The numerical results indicate that real 3D objects can be reconstructed clearly.

Acknowledgments. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(NRF-2017R1A2B4012096), and (NRF-2018R1D1A3B07044041)

References

1. J. H. Park, “Recent progress in computer-generated holography for three-dimensional scenes,” J. Inf. Disp, 18, 1(2017).
2. Y. Zhao, K. C. Kwon, M. U. Erdenebat, M.S. Islam, S. H. Jeon, and N. Kim, “Quality enhancement and GPU acceleration for a full-color holographic system using a relocated point cloud gridding method,” Appl. Opt. Vol. 57, 4253(2018)

A Fusion Architecture of CNN and Bi-directional RNN for Pornographic Video Detection

KwangHo Song[1,] , Yoo-Sung Kim[1]

[1] Department of Information and Communication Engineering
Inha University
Incheon 402-751, Korea
crossofjc@gmail.com, yskim@inha.ac.kr

Abstract. Automatic pornography detection is an important element to create a safe and clean content distribution circumstance in online space. In this paper, therefore, we propose the convolutional neural network (CNN) and Bi-directional Recurrent neural network(Bi-RNN) based pornography detection scheme. This model use a part of pre-trained CNN architecture, VGG-16, to CNN to extract spatial and static feature from each frame image and Bi-RNN to reflect temporal characteristic on video level such as motion and train overall network by end-to-end manner to classify the pornographic video. As a result, the average accuracy is 99.7% and this is a higher figure from at least 5% up to 17% more than other previous works.

Keywords: Pornography detection, Convolution Neural Network, Bi-directional Recurrent Neural Network

1 Introduction

In the recent situation that the amount of pornographic contents distributed by online [1], the demands about automatic filtering of pornographic contents on internet is also naturally increasing. Therefore the researches for automatic filtering of pornographic video contents have been proposed until a recent days, especially using deep learning [2-4].

In the research of [2-4], they utilized different architecture or fusion method depending on the modality that they importantly used to handle the 'Pornographic Video'. According to [2-4], three different modal such as image, video and motion can be extracted from the 'Video'. And by the modal that they used centrally, the video can be differently understood as a summation of individual images [2], intrinsic ensemble of frames [3] or the aggregation of multiple modal including the motion over time-series to enhance the effect of other modal [4]. So, although [2-4] commonly used pre-trained deep learning architecture as a feature extractor of frame in video, they used different detecting strategy such as feature-level fusion method[2], prediction ensemble by single[2] or multiple[3] detectors, which optimized for each modal, depending on the modality they used. And depending on the each way, there are some disadvantages such as difficulty in classifying certain type of data [2], impossibility of end-to-end learning [3] or overall performance degradation caused by a few piss-poor component detector [4].

Thus, we propose the fusion scheme consist of CNN and Bi-RNN for pornographic video detection. This scheme basically follows the argument of [4] that multi-modality is needed for the better classification, but doesn't need any prediction ensemble by multiple detectors or feature fusion. Instead, by end-to-end learning the static features, which extracted from frames by CNN, through bi-RNN, the network can learn both static and contextual features included in video.

2 Relate work

The definition about what is 'pornographic video' is the first thing to do to classify itself from another. In the past, the existence of local elements such as the appearance of certain parts of body or the intensity of nudity at frame and including such frame in video was an important factor to classifying pornographic video, but recent researches use more general definition to judge pornographic video regardless of the existence of such local elements [2-4].

In [2], they define the pornographic video is a summation of individual images which has sexual and risky scenes. So they fine-tune the various convolutional neural network such as AlexNet [5], GoogLeNet [6] to make pornographic image classifier and merge the softmax probabilities on image level into single decision result on video level by late fusion using average pooling. However, if the performer's appearance in the frames of pornographic video are not pornographic such as non-risky costume or, the other way, performer's appearance in the frames of non-pornographic video are pornographic such as feeding baby or beachside, performance of detector is difficult to guarantee because it is pornographic image detector, not video detector. On the contrary, [3] proposed pornographic video detection scheme using single video descriptor by video, which is created by feature fusion by average pooling of image descriptor extracted from each frame using a part of pre-trained VGG-16 [7], because they defined the pornographic video is an intrinsic ensemble of pornographic images and non-pornographic one related each other. Also the video descriptor is used to train support vector machine (SVM) classifier to discriminate which video is pornography. Therefore, because the classifier is separated from network used for feature extraction, it can't be trained by end to end manner.

Meanwhile, in [4], they proposed ensemble of multiple pornographic video detectors that each use different modal, which is image, video and motion included in video, because they think the definition of pornography should include not only stationary sexual or risky image, but also lewd motions performed over time on video. So for the detector of image and video, they used identical architecture with [2] and [3]. And for motion detector, they modified the architecture of [3] which use dual input to accommodate stacked motion frames of x and y direction extracted by optical-flow algorithm [8]. Using these three detectors, they filter the pornographic video contents from others by stacking ensemble method. Through this scheme, they can get improved accuracy of prediction than [2, 3], however a different optimal threshold should be assigned for each detector to be used in each step of stacking and if performance of any one of the component detector is bad, overall performance is decreased rapidly even if the others has fine performance.

Therefore in this paper, we propose the CNN and Bi-RNN based pornographic video detection scheme which can make single decision result on video level by itself. The architecture contain CNN to extract spatial and static feature from each frame image of video and Bi-RNN to reflect temporal characteristic on video level such as motion. Also because this scheme has an unsegmented architecture, it can be trained by end-to-end manner easily.

3 Pornographic video detection based on CNN and Bi-RNN architecture

In this section, we propose a pornography detection scheme based on CNN and Bi-RNN architecture. First of all, the meaning of 'pornographic video' that we want to find is an every video containing any single obscene frame image or lewd action performed on video. Thus, we adopt CNN to extract a useful features from every single frame, and adopt Bi-RNN to catch the visual contextual or temporal characteristics including motion information which spread along overall temporal flow of video. And the network architecture of the proposed detection scheme is suggested in Fig.1.

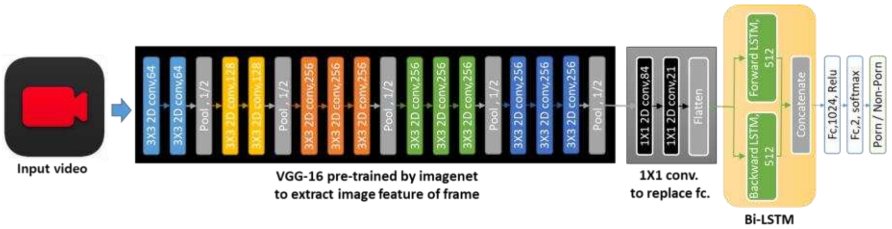


Fig. 1 network architecture of the proposed detection scheme

As in Fig. 1, the network is divided into up to four parts: 1. pre-trained VGG-16 [7] to extract feature matrix from frame image, 2. 1 by 1 convolution layers for replacing fully connected layers of VGG-16, 3. Bi-directional RNN layer using LSTM [9] cell to reflect the overall contextual characteristic of pornographic video including lewd motion, 4. Fully connected layers for classification.

Before explanation about network architecture, we need some pre-processing to modify the input video into a suitable form for network. So the one raw frame per second (1 fps) are extracted randomly from a raw input video data and they are revised into 224*224*3 size which is the same size of VGG-16 input layer and if the frame should be extended, then linear interpolation method is used. After then those modified input video is used for training the network and testing it.

Aforementioned, proposed scheme adopt a part of VGG-16, from input layer to pooling layer after 13th convolution layer, with pre-trained weight by imagenet as a feature extractor for randomly chosen frame images. It is one of the popular CNN architecture and it is easy to apply because of its simplicity. And also it has enough performances in terms of time and accuracy.

After then, 1 by 1 convolution layers are used to replace fully connected layers of VGG-16. While convolution operation using 1 by 1 kernel is identical with the operation of fully connected (fc) layer on mathematically, the number of parameter to be trained is smaller than fc layer. Therefore when the number of data available for fine-tuning is limited, 1 by 1 convolution layer can be used as a fine substitute because it can contribute to reduce the effort and resource for fine-tuning.

Through a series of layers, the frame images are transformed to feature matrices of each frame and they are transformed to the single feature vector on video level which contain overall contextual and temporal information by Recurrent neural network layer (RNN) using LSTM cell. By using RNN to make video level feature vector instead of using any other feature aggregation likewise [2] or [4], we can reasonably pour the characteristic of temporal flow into feature vector on video level than average pooling which used by [2,4]. Furthermore, since the frame at a particular time is highly influenced by both the preceding and the later scene on a video with a story and the characteristic of motion that often appeared in pornographic video is highly periodic and repetitive in any direction of video flow, Bi-RNN architecture, which can consider all the effects of bidirectional video flow, is an appropriate choice.

In the last of process, the feature vector on video level which generated by Bi-RNN is used to predict pornographic probability of video by fully connected layers and softmax activation function without result ensemble with any other detectors to make video level final decision. Accordingly, the worry about performance degradation by multiple detectors or the burden to choose optimal threshold for each of them is vanished.

4 Experiment

Before describing the experiment results, we utilized 'pornography-2k' dataset [10] which contain 1000 porn videos and 1000 non-porn videos, which classes in non-porn videos are variously mixed from easy sample to difficult one to classify like wrestling, person on beachside, feeding babies, to make the dataset that we used for training and test network. But, because each of videos in 'pornography-2k' have different running time, fps, and frame size, we randomly pick out some videos from there and split them into 10 seconds clip videos. As a result, we had created a video dataset with total of 6600 which consist of 3500 non-pornographic videos and 3100 pornographic videos. To use them for training and testing, they are divided into 3 part: 90% of data is used for training and 5 % is used for validation while training step and last remaining 5% data is used to test the trained network.

Based on these dataset, we performed some experiments to verify the performance of proposed scheme. The first experiment in Table 1 is conducted to find the best combination of pre-trained VGG-16 for image feature extraction and RNN architecture for contextual information generation.

Table 1. Performance evaluation by the combination of image feature extractor and RNN architecture

Extractor	RNN architecture	Validation acc.	Test acc.
VGG-16	Bi-direction	98.41	99.69
	Forward	98.25	98.18
	Backward	98.56	98.48

As a result of first experiment in Table 1, the maximum average accuracy was 99.69% by the combination of VGG-16 and Bi-directional RNN using LSTM cell. This result is a minimum about 5% higher than any other related works [2, 3]. And the second experiment in Table 2 is performed to compare the performance according to the multi-modality ensemble method in [4].

Table 2. Performance comparison with ensemble scheme

Method		Test acc.
Multi-modal (image/video/motion) ensemble	Stacking [4]	95.02
	Bagging (same as late fusion in [2])	97.14
Proposed scheme		99.69

As the result of second experiment in Table 2, proposed scheme also get the better performance than any other ensemble of multi-modal detectors which means proposed scheme more well reflect the contextual characteristic and motional information on video than other multi modal method which include its exclusive modal detector.

5 Conclusion

In this paper, we proposed the pornographic video detection scheme based on CNN and Bi-RNN. And we progress experiment to measure the performance of the proposed scheme. As a result, we can obtain 99.69% average accuracy and it is about 5% higher than the result of other previous related works. In the future, we try to produce the pornography detector based on dual stream architecture which can use not only visual contents but also acoustic contents to detect pornographic video.

References

1. Asia economy news, “The boom of pornography,prostitution related information in SNS and major potal… NAVER>Kakao”, http://www.asiae.co.kr/news/view.htm?idxno=2017022113542145899
2. M. Moustafa, “Applying deep learning to classify pornographic images and videos”, arXiv:1511.08899 [cs.CV]
3. KwangHo Song, and Yoo-Sung Kim, “Pornographic Video Detection Scheme using Video Descriptor generated by Deep Learning Architecture”, Proceeding of 4th International

Conference on “Emerging Trends In Academic Research” (ETAR- 2017), vol.4, 2017, pp. 59-65

4. KwangHo Song, and Yoo-Sung Kim, “Pornographic Video Detection Scheme Using Multimodal Features”, Journal of Engineering and Applied Sciences vol. 13, No. 5, 2018, pp. 1174-1182

5.Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Proceeding of the Advances in Neural Information Processing Systems 25(NIPS 2012), pp. 1097-1105

6. Christian Szegedy, Wei Liu, Yangqing Jia Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, “Going Deeper with Convolutions”, arXiv:1409.4842 [cs.CV]

7. Karen Simonyan, Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, arXiv:1409.1556 [cs.CV]

8. Farneback, G. “Two-Frame Motion Estimation Based on Polynomial Expansion”, In: Image Analysis, Bigun, J. and T. Gustavsson (Eds.), Springer, Bernlin, Germany, ISBN:978-3-540-40601-3, 2003, pp.363.

9. S. Hochreiter and J. Schmidhuber. “Long short-term memory”, Neural computation, Vol.9, No.8, 1997, pp.1735–1780

10. Moreira, D., S. Avila, M. Perez, D. Moraes and V. Testoni et al., Pornography classification: The hidden clues in video space-time. Forensic Sci. Int., Vol. 268, 2016 pp.46-61

Program Networks at Run-time for Bulk Data Transfer using AmoebaNet

Syed Asif Raza Shah[1], Seo-Young Noh[2*]

[1] Department of Computer Science,
Sukkur Institute of Business Administration University(SIBAU),
Pakistan. asif.shah@iba-suk.edu.pk
[2] Department of Computer Science, Chungbuk National University
Republic of Korea. rsyoung@cbnu.ac.kr

Abstract. Data transfer is an essential function for scientific discoveries, particularly within big data science. Although significant improvements have been made in the area of bulk data transfer, current available tools and services are not successfully addressing the high-performance and time-constraint challenges of data transfer for extreme-scale science applications due to disjoint end-to-end data transfer loops and cross-interference between data transfers. In this paper, we have prototyped AmoebaNet's SDN-enabled network service that allows application (e.g. BigData Express) to program networks at run-time for bulk data transfers. We presented preliminary results of AmoebaNet service.

Keywords: SDN, OpenFlow, Network as a Service, Bulk data transfer, QoS

1 Introduction

The emergence of distributed, extreme-scale science applications is generating significant challenges in data transfer. Data transfer is now an essential function for scientific discoveries, particularly within big data environments. To date, there are several data transfer tools (such as GridFTP [1] and BBCP [2]) and services (such as the PhEDEx high-throughput data transfer management system, the LIGO Data Replicator, and Globus Online [3]) have been developed to support bulk data transfer. Number of advanced data transfer features, such as transfer resumption, partial transfer, third-party transfer, and security, have been implemented in these tools and services. Although significant improvements have been made in bulk data transfer, currently available data transfer tools and services will not be able to successfully meet these challenges due to the lack of data-transfer-centric approach, the lack of effective mechanisms to minimize cross-interference between data transfers and lack of QoS.

In this paper, we have introduced a Software-Defined Network (SDN) [4] enabled service known as AmoebaNet [5] that allows application to program networks at run-time for bulk data transfers.

[*] Corresponding Author

2 Related Works

2.1 Overview of AmoebaNet

AmoebaNet project is currently ongoing evaluation and enhancement. There are several design goals for AmoebaNet. (1) "Network as a Service" is our primary goal. AmoebaNet must allow application to program network at run-time for optimum performance. (2) QoS guarantee. AmoebaNet must provide QoS guarantee for priority traffic. (3) Wide applicability. AmoebaNet was originally designed to support BigData Express (BDE) [6] project. However, AmoebaNet has wide applicability to support a wide range of applications[7].

AmoebaNet is an enterprise network service. Along with WAN Connection Service such as ESNet OSCARS [8, 9], it can fast provision end-to-end network paths with guaranteed QoS across domains. We implemented AmoebaNet using Java upon the ONOS [10] platform. It has 3K+ lines of code. The latest AmoebaNet is released at: http://bigdataexpress.fnal.gov/Releases.html.

2.2 BigData Express and AmoebaNet service

A major goal of BigData Express is to apply AmoebaNet's SDN-enabled network service to provide "Application-aware" network and facilitate data transfer. BigData Express will typically run in a large data center, such as the DOE Leadership Computing Facilities. Such a site will typically feature a dedicated cluster of high-performance DTNs, an SDN-based BigData-Express LAN, and a large-scale storage system. For data transfer jobs a logically centralized BigData Express scheduler will coordinate all activities at each BigData Express site. This BigData Express scheduler will manage and schedule local resources (DTNs, storage, and the BigData Express LAN) through agents (DTN agents, storage agents, and ***AmoebaNet***). ***AmoebaNet*** will keep track of the BigData Express LAN topology and traffic status with the aid of SDN controllers. As requested by the BigData Express scheduler, it will program network at run-time for data transfer tasks.

3 Evaluation and Results

We evaluated AmoebaNet service with a data transfer test case. In this experiment, we evaluated AmoebaNet in a cross pacific domain and used BigData Express to demonstrate its key features and capabilities. AmoebaNet provided APIs to BDE to program the underlying campus network for provisioning on-demand end-to-end QoS guaranteed paths.

Our experimental topology consists of two independently managed scientific computing facilities – Fermi National Accelerator Laboratory(FNAL) in the U.S.A site and Korea Institute of Science and Technology Information(KISTI) site in Korea, with a dedicated layer-2 WAN circuit that connects both sites. The maximum available end-to-end bandwidth between FNAL and KISTI was 10Gbps. In this topology the FNAL site was logically divided into two different sites which consist of the separate AmoebaNet-based SDN controllers, SDN switches and DTNs. Sites configurations as follow:

FNAL sites:

Site 1: DTNs: BDE1, BDE2, BDE3.

SDN switches: Pica8 P5101
(running PicaOS)
An AmoebaNet-based SDN controller

Site 2: DTNs: BDE4, BDE-HP5.

SDN switches: Pica8 P3930.
(running PicaOS)
An AmoebaNet-based SDN controller

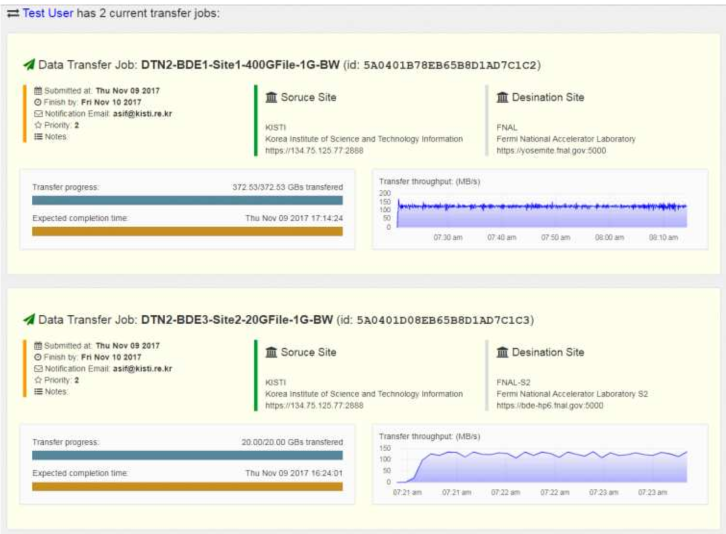
KISTI site:

DTNs: DTN2, DTN3, and DTN4.
SDN switches: HP Z91000.
(running PicaOS)
An AmoebaNet-based SDN controller

To evaluate AmoebaNet functionality we configured AmoebaNet and all related software of BDE at FNAL and KISTI sites. The data transfer service is accessible for users using BDE web portal: https://yosemite.fnal.gov:5000 (accessible URL at FNAL site), or https://134.75.125.77:2888/ (accessible URL at KISTI site), respectively. In the evaluation, two parallel data transfer tasks were submitted at https://134.75.125.77:2888 (BDE web portal at KISTI):

Task 1: 1Gpbs dedicated bandwidth for QoS-guaranteed end-to-end path and transferred 372.5GB data set from DTN2 at KISTI to BDE1 at first logical site of Fermilab (i.e. FNAL).

Task 2: 1Gbps dedicated bandwidth for QoS-guaranteed end-to-end path and transferred 20GB data set from DTN2 at KISTI to BDE3 at first logical site of Fermilab (i.e. FNAL).



Test User has 2 current transfer jobs:

Data Transfer Job: **DTN2-BDE1-Site1-400GFile-1G-BW** (id: 5A0401B78EB65B8D1AD7C1C2)

Submitted at: **Thu Nov 09 2017**
Finish by: **Fri Nov 10 2017**
Notification Email: **asif@kisti.re.kr**
Priority: **2**
Notes:

Soruce Site
KISTI
Korea Institute of Science and Technology Information
https://134.75.125.77:2888

Desination Site
FNAL
Fermi National Accelerator Laboratory
https://yosemite.fnal.gov:5000

Transfer progress: 372.53/372.53 GBs transfered

Expected completion time: Thu Nov 09 2017 17:14:24

Data Transfer Job: **DTN2-BDE3-Site2-20GFile-1G-BW** (id: 5A0401D08EB65B8D1AD7C1C3)

Submitted at: **Thu Nov 09 2017**
Finish by: **Fri Nov 10 2017**
Notification Email: **asif@kisti.re.kr**
Priority: **2**
Notes:

Soruce Site
KISTI
Korea Institute of Science and Technology Information
https://134.75.125.77:2888

Desination Site
FNAL-S2
Fermi National Accelerator Laboratory S2
https://bde-hp6.fnal.gov:5000

Transfer progress: 20.00/20.00 GBs transfered

Expected completion time: Thu Nov 09 2017 16:24:01

Figure 1: Evaluation results of parallel submitted tasks

Both tasks were submitted in parallel and started data transfer job at the same time. When both tasks submitted using BDE web portal after that, the BDE software starts communication with AmoebaNet and send queries for available bandwidth of each local sites. Once the available bandwidth satisfies the requirements, then BDE software sends JSON formatted command using APIs to AmoebaNet for reservation of bandwidth at both sites.

The evaluation results are illustrated in Figure 1. The task 1 successfully completed its data transfer job from KISTI (DTN1) to FNAL (BDE1) at the transfer rate of approx. ~130MB/s which is equal to ~1Gbps. On the other hand, the task 2 also completed its jobs from KISTI (DTN1) to FNAL S2 (DTN4) and the data transfer rate was approximately ~130MB/s which is also equal to ~1Gbps.

4 Conclusion

In this paper we have shown how to program the network at run time for different data transfer jobs using AmoebaNet and BigData Express application. An initial version of AmoebaNet has been tested in scientific SDN testbeds. The evaluation results show that AmoebaNet complements existing network paradigm and provide end-to-end QoS network services for data science.

Acknowledgments. This work has been done while the authors were participating in the joint project between Fermi National Accelerator Laboratory and Korea Institute of Science and Technology Information. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2008-00458).

References

[1] B. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu and I. Foster, “The Globus Striped GridFTP Framework and Server,” SC'2005, 2005.

[2] BBCP, http://www.slac.stanford.edu/~abh/bbcp/

[3] https://www.globus.org/

[4] McKeown, Nick. “Software-defined networking.” INFOCOM keynote talk 17.2 (2009): 30-32.

[5] S.Asif Raza, et al. "*AmoebaNet: An SDN-enabled network service for big data science*." Journal of Network and Computer Applications 119 (2018): 70-82.

[6] http://bigdataexpress.fnal.gov/

[7] Vinoski, Steve. "Advanced message queuing protocol." IEEE Internet Computing 10.6 (2006).

[8] Guok, Chin P., et al. "A user driven dynamic circuit network implementation." GLOBECOM Workshops, 2008 IEEE. IEEE, 2008.

[9] Roberts, Guy, et al. "Nsi connection service v2. 0." GFD. 212 (2014): 1-119.

[10] Shenker, Scott, et al. “The future of networking, and the past of protocols,” Open Networking Summit (2011).

Improved U-Net for Lung Tumor Segmentation using Attention mechanism and Tversky loss

Trinh Le Ba Khanh[1] , Do Truong Dong[1] , Hyung-Jeong Yang*[1]

[1] Dept of Electronics and Computer Engineering, Chonnam National University, Gwangju,
South Korea
51304375@hcmut.edu.vn, truongdong24696@gmail.com, hjyang@jnu.ac.kr,

Abstract. Lung cancer is one of the most leading causes of death worldwide. Early detection of cancer in computed tomography (CT) image is a key to beating cancer. Recently, convolutional neural networks have been applied to various medical image analysis tasks. In this paper, we proposed a deep learning based approach for automatically segmenting the lung tumor by utilizing U-net on CT images. Due to the small fraction of the tumor in the whole lung image, we try to balance the foreground and background by using the Tversky loss as well as focus on target region by Attention mechanism, which will improve the accuracy in segmentation. The proposed model get average dice coefficient of 61.81 on lung tumor dataset provided by VIP-CUP 2018.

Keywords: Lung Cancer Tumor Segmentation, U-net, CT Scans

1 Introduction

Lung cancer is one of the widespread disease and the leading causes of death worldwide. As reported in [1], lung cancer, causing 1.3 million deaths annually, is one of the leading causes of cancer death worldwide. In order to decrease the number of death, detection and treatment at an early stage for potential patients are crucial. Especially, lung tumor cancer segmentation is worth of attention. Due to the development of computed tomography (CT) imaging, it was adopted as a standard modality for assessing lung cancer.

Recently, Convolutional neural networks (CNNs) have produced state-of-the-art results for image classification and segmentation. CNNs have been applied to various medical image analysis tasks and have shown great potential for medical applications such as brain tumor segmentation [2], liver tumor segmentation [3], pancreas segmentation [4] and in computer-aided diagnostic applications [5]. State-of-the-art computer vision techniques and deep learning has given more opportunity for beating cancer.

In this paper, our goal is to segment the lung cancer tumor given the 2D data and the corresponding segmentation results for each patient. We proposed a model for automatically segmenting the lung tumor by utilizing U-Net on CT data. The proposed model has gotten the promising result on segmentation lung tumor task.

The remainder of the paper is organized as follows. Section 2 reviews some related researches. Section 3 describes the proposed system. Section 4 shows the data information and experimental results of the proposed model and the last one is conclusion in Section 5.

2 Related Works

With the availability of large data of medical image data as well as technological improvements, deep learning based approaches have taken the lead in most medical image analysis tasks. A common task in medical image analysis is the ability to detect and segment pathological regions that occupy in the image. In recent research, convolutional neural networks (CNNs) have been successfully applied to automatically segment 2D and 3D biological data [6]. High representation power, fast inference and filter sharing properties have made CNNs the one standard for image segmentation [7]. One of the most popular networks for segmentation is encoder-decoder convolutional neural network architecture called U-Net [8]. Due to its multi-scale skip connections and learnable up-convolution layer, U-net network becomes the standard structure for image segmentation.

Regarding tumor segmentation, tumor regions typically occupy a very small fraction of the full image, one of the efforts to address small ROI segmentation such as attention gated networks [9]. CNNs with attention gates (AGs) automatically learn to focus on the target region without additional supervision and can be trained end to end similar to the training of a FCN model [7]. At test time, these gates generate soft region proposals to highlight salient features useful for a specific task.

In this research, to address the issues of data imbalance and improve the performance, we combine attention gated U-Net with a Tversky loss function [10], suited for small region segmentation. The performance was shown in dataset provided by VIP-CUP 2018 [11] with dice scores of 61.81, outperform the 5 point gap when compared to the standard U-Net.

3 Proposed Method

In this section, we first describe the proposed structure for segmenting. We then present the Tversky loss for training model. Finally, implementation details are provided.

The U-Net model [8] has shown outstanding performance on medical segmentation task. To increase the information of feature extraction as well as reduce the training time, we replaced the backbone of U-Net by using the pre-train model on ImageNet dataset, thus U-Net can achieve good performance with a limited amount of training data. The backbone of U-Net used Resnet50 [12] for all experiment in this research. The U-Net structure used in the paper is shown in Fig. 1

On the other hand, at the deep stage of encoding, the network has the rich semantic information, but spatial details tend to get lost, which make difficult to detect and segment small object like tumor. We use soft attention gates (AGs) to address this issue. AGs can automatically focus on relevant region from low-level feature maps in the encoder and propagate it to the decoder, as depicted in Figure 2.

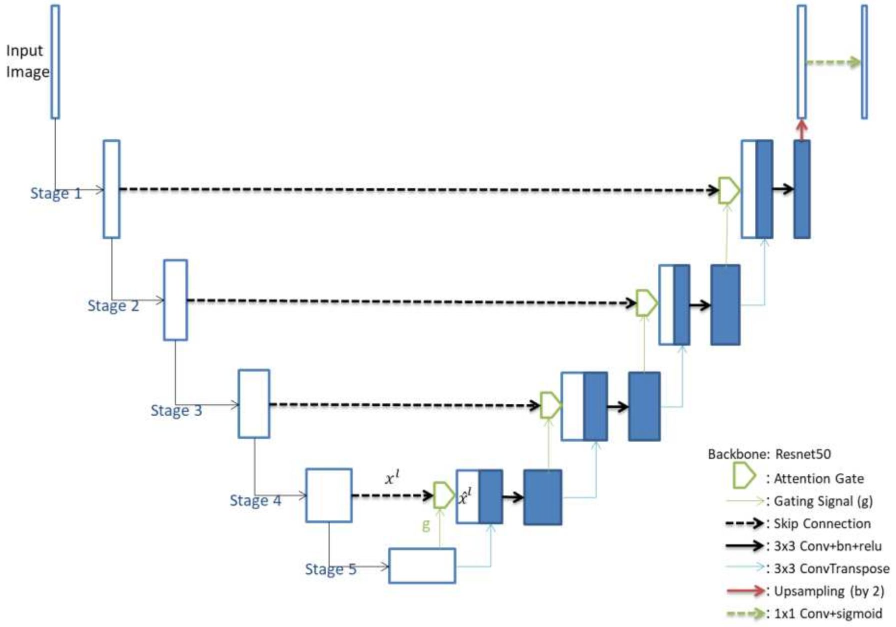


Fig. 1. Proposed U-Net architecture with Attention Gate.

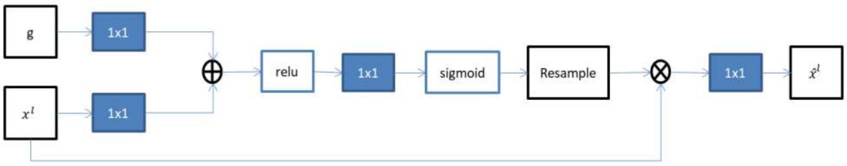


Fig. 2. Schematic of the Attention Gate.

The limitation of the Dice loss [13] is that it equally weighs of false positive and false negative. Since the task is a lung tumor segmentation, the data distribution of foreground and background classes is a high imbalance, we then apply the Tversky loss introduced by [10] in our method. Thus, the Tversky loss can add weight to false positive and false negative, was written as follow:

$$T(\alpha,\beta) = \frac{\sum_{i=1}^{N} p_{1i} g_{1i}}{\sum_{i=1}^{N} p_{1i} g_{1i} + \alpha \sum_{i=1}^{N} p_{0i} g_{1i} + \beta \sum_{i=1}^{N} p_{1i} g_{0i}} . \quad (1)$$

where in the output, the $\mathrm{p}_{1\mathrm{i}}$ is the probability of pixel *i* be a tumor and $\mathrm{p}_{0\mathrm{i}}$ is the probability of pixel *i* be a non-tumor. Also, $\mathrm{g}_{1\mathrm{i}}$ is 1 for a tumor pixel and 0 for a non-tumor pixel and vice versa for the $\mathrm{g}_{0\mathrm{i}}$.

The proposed model is implemented in Keras framework with Tensorflow backend. The input and output have the sizes 512x512x3 and 512x512x1, respectively. We use SGD optimizer for training with the learning rate of 0.001 and batch size of 4. We choose α=0.8 and β=0.2 in Equation 1 for all experiments. Our models were trained using Tensorflow-GPU with NVIDIA GTX 1080 Ti GPU.

4 Experimental Result

We have used the Dataset provided by the 2018 IEEE VIP CUP. The training set consists of CT scans of 260 patients while the validation set comprises 40 patients. Both the training and the validation are manually annotated by a radiation oncologist. The test set has 40 patients without manual annotation. We use the training set for training purpose and validation set for evaluating. We adopt the dice coefficient as an evaluation metric to evaluate the effectiveness of the proposed method. The results are presented in Table 1. The proposed model got a much higher dice coefficient. The Attention mechanism and Tversky loss have contributed the important roles to improve the accuracy of small ROI segmentation. Figure 3 show some examples of the segmentation result.

Table 1. Segmentation result on VIP CUP 2018 data.

Model	Dice Score
[14]	59.20
U-Net + Dice loss	57.28
U-Net + Tversky loss	59.94
Attention U-Net + Tversky loss (proposed)	**61.81**

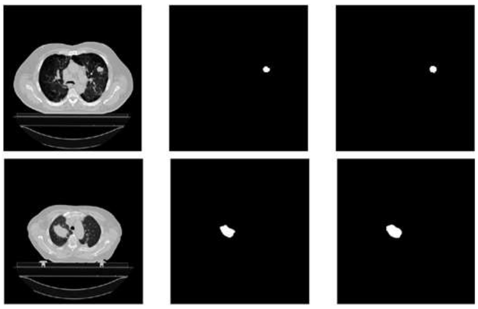


Fig. 3. Examples of segmentation result (Left: Image, Middle: Ground Truth, Right: Predicted).

5 Conclusion

In this paper, the proposed segmentation model was seen to perform better by almost 5 point gap in terms of dice coefficient compared to the standard U-Net model. The proposed model was able to automatically focus on the small region of the image by taking advantage of the attention mechanism. Additionally, our experiments demonstrate the importance of the choice of loss function when dealing with highly imbalanced problems, especially for small region segmentation such as lung tumor segmentation. In further works, we planned to extend the model with deeper stacks of slices to get more useful spatial information.

Acknowledgments
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP)(NRF-2017R1A2B4011409)

References

1. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2015. CA Cancer J. Clin.65(1), 5--29 (2015)
2. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al.: Brain Tumor Segmentation with Deep Neural Networks. Med. Img. Ana. 35, 18--31 (2017)
3. Li, W., Jia, F., Hu, Q.: Automatic Segmentation of Liver Tumor in CT Images with Deep Convolutional Neural Networks. J. Comp. Com. 3, 146--151 (2015)
4. Roth, H.R., Farag, A., Lu, L., Turkbey, E.B., Summers, R.M.: Deep Convolutional Networks for Pancreas Segmentation in CT Imaging. Proceedings of Society of Photographic Instrumentation Engineers 9413, Medical Imaging 2015: Image Processing, 94131G (2015)
5. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al.: Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Transactions on Medical Imaging. 35, 1285--1298 (2016)
6. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 240--248 (2017)
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CVPR, pp. 3431--3440 (2015)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. MICCAI, pp. 234--241 (2015)
9. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., et al.: Attention U-Net: Learning Where to Look for the Pancreas. MIDL (2018)
10. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. MLMI, pp. 379-387 (2017)
11. 2018 IEEE Video and Image Processing Cup (VIP-Cup), https://users.encs.concordia.ca/~i-sip/2018VIP-Cup/index.html.

12. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. CVPR (2016)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. 3DV (2016)
14. Zhang, R., Guan, Z., Lai, S.C., Xiao, J., Lam, K.M.: Deep Neural Networks for Lung Cancer Tumor Region Segmentation. IWAIT (2019)

Scalable Data Science and Machine Learning Algorithm for Gene Prediction

Oluwafemi A. Sarumi[1,2][0000−0001−6463−1029] and
Carson K. Leung[2][0000−0002−7541−9127]

[1] The Federal University of Technology - Akure (FUTA), Ondo State, Nigeria
oasarumi@futa.edu.ng
[2] University of Manitoba, Winnipeg, MB, Canada
kleung@cs.umanitoba.ca

Abstract. Recent technological advances and scientific discoveries have revolutionized the current era of genomics. The use of next-generation sequencing (NGS) technologies has tremendously reduced the sequencing time and given rise to the production and collection of high volumes of genomic datasets. Predicting protein-coding genes from these copious genomic datasets is significant for the synthesis of protein and the understating of the regulatory function of the non-coding region. Over the past few years, researchers have developed methods for finding protein-coding genes in the genome of organisms. Notwithstanding, the recent data explosion in genomics accentuates the need for more efficient algorithms for gene prediction. In this paper, we propose a scalable naïve Bayes-based machine learning algorithm that is deployed over a cluster of Apache Spark framework for efficient prediction of genes in the genome of eukaryotic organisms. Evaluation results on discovering the protein-coding genes from the human genome chromosome GRCh37 show that our algorithm led to high sensitivity, specificity and accuracy.

Keywords: Genomics · Machine learning · Data science · Gene prediction · Protein synthesis · Big data · Bioinformatics · Apache Spark · Data analytics · Big data applications and services

1 Introduction and Related Works

Over the past few years, technological advances and scientific discoveries have revolutionized bioinformatics. The emergence of high-throughput next-generation sequencing (NGS) technologies—such as Illumina HiSeq X and Illumina Genome Analyzer—has tremendously reduced the sequencing time and given rise to the production and collection of high volumes of omics data (e.g., genomics, proteomics, transcriptomics, metagenomics). In genomics, petabytes of organisms (e.g., eukaryotes and prokaryotes) complete sequenced genomes are currently available in several public repositories. This avalanche of genomic datasets introduced new challenges for researchers, and demanded for improved computational approaches for some bioinformatics tasks such as gene prediction [1], motifs discovery [2, 3], sequence alignment [4], and sequence assembly [5].

Gene prediction involves the process of locating the regions that encode the protein-coding genes and other functional elements in genomic datasets. In eukaryotic organisms [6], gene prediction is a rigorous task due to (a) the problem of inconsistency between genes and (b) a very small amount of genes containing in most of the genomes. For example, protein-coding gene is less than 5% of the entire human genome (cf. non-coding elements [7,8]). In addition, further complexities arise from the long distance between the exons (coding region), limited knowledge of the promoters, and the existence of an alternative splicing site after the transcription [9]. Moreover, the introns (i.e., non-coding regions) that separate the exons and the splicing sites (i.e., a region that divides the exons from the introns) are difficult to identify due to its wide distance and the undefined length.

In prokaryotic organisms [10], the protein-coding region can be identified in a contiguous sequence [11] known as an *open reading frame* (*OFR*) without the interruption of introns. Hence, identifying exons in genomes of prokaryotic organisms is considered as a less rigorous task.

Protein-coding regions can be identified in a contiguous seq [12,13]. The first method is the ab initio gene prediction [14] that makes use of gene structure as a template to detect genes. This is done by logically examining and separating signal sensors [15] and distinct biological (content sensors) [16] patterns as well as being able to distinguish gene regions in a single input sequence. Several ab initio gene prediction methods [17–20] have been proposed in the literature by researchers. Ab initio gene finding approaches have been found limited [15,21] due to the fact that little knowledge of gene structures is available, especially for new sequencing genomes. In addition, it is difficult to detect periodicity and other known content properties of protein-coding genes.

The second approach is based on sequence similarity searches [22,23]. Sequence similarity search finds similarity in gene sequences between expressed sequence tags (ESTs), proteins, or other genomes to the input genome. This approach assumes that the coding regions are more conserved evolutionarily [24] than non-coding regions. Several authors [25–28] have proposed different gene prediction methods based on the sequence similarity search. EST-based sequence similarity usually has bottlenecks [21,29] due to the fact that ESTs only correspond to small portions of the gene sequence, which makes it difficult to predict the complete gene structure of a given region.

A recent approach to gene prediction is based on machine learning algorithms and data mining techniques [30,31]. A major advantage of ML methods is its ability to automatically identify patterns in data [32]. This is highly important when the expert knowledge is incomplete or inaccurate when the amount of available data is too large to be handled manually, or when there are exceptions to the general cases. Several researchers [33–36] have reported a machine learning approach for gene prediction in genomes of organisms.The major limitations of the previously reported ML algorithms for gene predictions are: (i) they cannot scale to handle the current large volumes of genomic datasets, and (ii) they were highly restricted to a type of organism or dataset. Hence, there is a need for more efficient, scalable ML algorithms that can improve with experience for predicting genes from big volumes of a genomic dataset. Our ***key contribution*** of this paper is our scalable naïve Bayes machine algorithm for identification of protein-coding regions in large volumes of eukaryotic organisms genome.

The remainder of this paper is organized as follows. The next section presents the background. Section 3 discusses the data preprocessing actions. Section 4 discusses our scalable NBML algorithm. Experimental results are shown in Section 5, and conclusions are given in Section 6.

2 naïve Bayes Algorithm and Apache Spark Framework

ML algorithms are used to develop models that learn from experience and discover novel patterns in a dataset. Generally, they are divided into two categories as supervised [37] and unsupervised learning [38]. naïve Bayes algorithms [39, 40] are supervised, and probabilistic classifiers developed on Bayes theorem. They are based on the assumption that inclusion or exclusion of a particular feature in the model is independent on the inclusion or exclusion of any other feature and their contributions towards probability is independent of each other. A major advantage of NB classifier is that a small amount of training data are required for estimating the parameters for classification. Given a set of items, each of which belongs to a known class C, and each of which has a known vector of variables v, the goal of NB is to construct a rule which will allow us to assign future items to a class in C, given only the vectors of variables describing the future items. NB algorithm can be scaled for big data analytics and applications by deploying it on Apache Spark framework.

Over the past few years, Apache Spark has attracted a lot of attention in the research community for processing high volumes of big data on a distributed system. Spark application runs as independent sets of processes on a cluster, coordinated by the SparkContext object in the driver program on the master node. The driver program connects to one or more worker nodes through the cluster manager. The driver program defines various transformations and invokes consequence actions on worker nodes. Spark actions are executed through a set of stages, separated by distributed operations and are made possible by the use of broadcast variables that allows programmers to keep a read-only variable cached on each machine rather than shipping a copy of it with the task.

The advantages of using Apache Spark framework for big data processing include the following [3]:

1. The driver program serves as a resource distributor and a result; collector
2. lost partitions can easily be reconfigured without the loss information;
3. Spark stores intermediate results in memory instead of disk;
4. it has rich support of various system workloads such as batch processing, iterative, interactive processes, machine learning, and graph processing;
5. worker nodes serve as computing units handling sub-tasks.

3 Data Preprocessing and Munging

The human genome (Homo sapiens) assemblies GRCh37 patch 13(hg19) and GRCh38 patch 10 (hg38) were obtained from the Ensembl data repository—www.ensembl.org. The human genome assemblies GRCh37 patch 13 is of size 3.2Gb and contains 104,763 protein-coding sequences (PCS) and 24,513 non-coding sequences (NCS). The genome GRCh38 patch 10 is of size 3.4 Gb and consists of 102,915 protein-coding sequences and 28,321 non-coding sequences. Short sequences were filtered from the human genome GRCh37 and GRCh38. After filtering, we obtained a sum of 94,830 protein-coding sequences and 24,266 non-coding sequences from the GRCh37 genome and 92,716 protein-coding sequences and 28,024 non-coding sequences from the GRCh38 genome. We selected 20,000 PCS and 20,000 NCS from the GRCh37 and 22,000 PCS and 22,000 NCS from the GRCh38 for the training and testing of our model. We converted our dataset into set codons. Codons are a group of nucleotides that specifies one amino acid. Furthermore, we discretized and labeled our dataset to format acceptable for the training. All the PCS were labeled as 1 and the NCS labeled as 0.

4 Our Scalable NBML Algorithm

Given that there are k possible classes of sequence $C = \{c_1, c_2, \ldots, c_k\}$ for a genome sequence $S = \{s_1, s_2, \ldots, s_n\}$, if $T = \{t_1, \ldots t_m\}$ be the set of unique codons that appears at least once in the genome sequence S. Then, the probability of a genome sequence s being in class c can be computed using the Bayes rules as shown in Eq. (1):

$$P(c|s) = \frac{P(c)P(s|c)}{P(s)} \tag{1}$$

where $P(s)$ is a constant for the known genome sequence size. Also, in Bayesian statistics, $P(s)$ is not often calculated for maximum a posteriori estimation problems. Hence, with NB, we can safely say that each codon t_j in the sequence occurs independently given the class c. Thus, Eq. (1) can be written as follows:

$$P(c|s) \propto P(c) \prod_{j=1}^{n_s} [P(t_j|c)]^{f_j} \tag{2}$$

where n_s is the number of unique codons in the genome sequence s and f_j is the frequency of each codon t_j. And, Eq. (2) becomes:

$$\log P(c|s) \propto \log P(c) + \sum_{j=1}^{n_s} [f_j \log P(t_j|c)] \tag{3}$$

If $c*$ defines the class of genome sequence that maximizes $\log P(c|s)$ in Eq. (3), then $c*$ can be written as follows:

$$c^* = \text{argmax}_{c \in C} \left\{ \log P(c) + \sum_{j=1}^{n_s} [f_j \log P(t_j|c)] \right\} \tag{4}$$

Thus, with NB classifiers, $P(c)$ and $P(t_j|c)$ can be estimated as shown in Eqs. (5) and (6), respectively:

$$\widehat{P}(c) = \frac{M_c}{M} \tag{5}$$

$$\widehat{P}(t_j|c) = \frac{M_{tj}}{\sum_{ti \in T} M_{ti}} \tag{6}$$

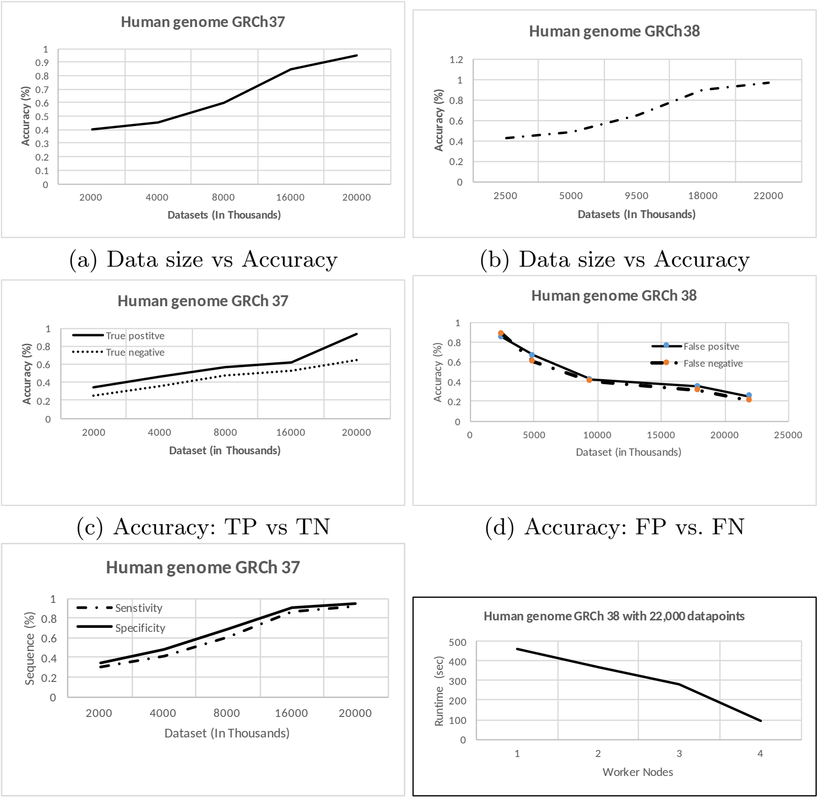
where M is the total number of genome sequence, M_c is the number of sequence in class c and M_{ti} is the frequency of a codon t_i in a class. The genome sequence is classified in to two classes with class 1 for the protein-coding genes and class 0 for the non-coding genes. Also, $P(c)$ from Eq. (5) is a constant since the number of protein-coding sequence is the same as the non-coding sequence. $P(t_j|c)$ is the frequency of codon t_j in all sequence in c. Thus, an overview of our Spark-based scalable NBML algorithm for identifying protein-coding genes in genomes is given as follows:

- submit the discretized and labelled data through the Master node.
- parallelized and slit the input data on all the worker nodes
- performing the training of the data on the workers coordinated through the master node
- aggregate the results of the training from the workers on the master node (the predictive model)
- use the model on the test data
- save model

5 Experimental Evaluation

To evaluate our proposed algorithm for predicting protein-coding genes from eukaryotic organism genome, 80% of the sequence from the GRCh37 and GRCh38 dataset were used as the training set and 20% used as the testing set. Our algorithm was implemented in Python on a standalone cluster of Apache Spark version 2.4.0. Our cluster runs on five machines (i.e., one master and four workers). The master node has (a) a processor of 3.33x4 GHz, (b) two dual Intel core i5 with 16 GB of RAM. Each worker node has a processor of 4.2 GHz, 56 cores with 125 GB of RAM. We also configured our Spark framework on Ubuntu-18.10 and 64-bit operating system.

We quantify the performance of our algorithm using the standard performance metrics. Sensitivity measures the proportion of the coding sequence in the genome that is correctly predicted as the coding sequence. Specificity measures the proportion of the non-coding sequence in the genome that is correctly predicted as a non-coding sequence. Accuracy measures the overall correct predictions in the genome. Figs. 1(a) and 1(b) respectively, show that the accuracy of our algorithm increases as the data size increases. Each accuracy point is calculated as the average of six runs of the algorithm. This implies that our algorithm performs better with larger datasets and our model becomes relatively stable from 19000 data points.



(a) Data size vs Accuracy

(b) Data size vs Accuracy

(c) Accuracy: TP vs TN

(d) Accuracy: FP vs. FN

(e) Sensitivity vs. Specificity

(f) Worker Nodes vs. Runtime

Fig. 1. Evaluation results

Furthermore, Fig. 1(c) shows that the true positives (TP) and true negatives (TN) increases as the size of the dataset increases while Fig. 1(d) shows that false positives (FP) and false negatives (FN) reduces as the size of the dataset increases. TP is set predicted as gene sequences that are known genes, FP is set predicted as gene sequences that are non-known genes, FN is set predicted as non-gene sequences that are known genes, and TN are set predicted as non-gene sequences that are non-known genes. Hence, it is logical that Fig. 1(e) shows that sensitivity and specificity increases as the data size increases. Also, we demonstrated the scalability of our algorithm in Fig. 1(f). It shows that the runtime of our algorithm reduces as the worker nodes on the cluster increases.

6 Conclusions

High volumes of a wide variety of valuable data can be easily collected and generated from a broad range of data sources of different veracities at a high velocity. In genomics, the use of next-generation sequencings (NGS) technologies such as Illumina HiSeq X and Illumina Genome Analyzer has tremendously reduced the sequencing time and given rise petabytes of eukaryotic and prokaryotic organisms complete sequenced genomes available in several public repositories. These copious genomic datasets introduced new challenges for researchers and the demands for improved computational approaches and methods for some bioinformatics task such as gene prediction motifs, discovery sequence alignment, and genome assembly. In this paper, we presented a scalable naïve Bayes ML algorithm deployed over a cluster of Apache Spark framework for efficient prediction of genes in the genome of eukaryotic organisms. Specifically, our algorithm was evaluated with two datasets, human genome GRCh 37 and GRCh 38. Experimental results show the performance of our algorithm in terms of accuracy, sensitivity, specificity, and scalability.

Acknowledgements

This project is partially supported by (i) The Association of Commonwealth Universities (ACU) in the form of the University of Manitoba Fellowship, (ii) The Natural Sciences and Engineering Research Council of Canada (NSERC), as well as (iii) University of Manitoba.

References

1. Alioto, T.: Gene prediction, In: Anisimova M. (eds) Evolutionary Genomics. MIMB, vol. 855, pp. 175-201. Humana Press, Totowa, NJ (2012) doi:10.1007/978-1-61779-582-4_6
2. Jiang, F., Leung, C.K., Sarumi, O.A., Zhang, C.Y.: Mining sequential patterns from uncertain big DNA in the Spark framework. In: IEEE BIBM, pp. 874-88 (2016)
3. Sarumi, O.A., Leung C.K., Adetunmbi, O.A.: Spark-based data analytics of sequence motifs in large omics data. Procedia Computer Science 126, 596-605 (2018)
4. Duan, X., Zhao, K., Liu, W.: HiPGA: a high performance genome assembler for short read sequence data, In: IEEE IPDPS 2014 Workshops, pp. 576-584 (2014)
5. Ekblom, R., Wolf, J.B.: A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications 7(9), 1026-1042 (2014) doi:10.1111/eva.12178

6. Gross, T., Faull, J., Ketteridge, S., Springham, D.: Eukaryotic microorganisms. In: Introductory Microbiology, pp. 241-286. Springer, Boston (1995)
7. She, R., Chu, J.S., Wang, K., Chen, N.: Fast and accurate gene prediction by decision tree classification. In: SIAM DM 2010, pp. 790-801 (2010) doi:10.1137/1.9781611972801.69
8. Do, J.H., Choi, D.K.: Computational approaches to gene prediction Journal of Microbiology 44(2),137-144 (2006)
9. Martins, P.V.L.: Gene prediction using deep learning. Master's dissertation, University of Porto (2018)
10. Margulis, L.: The classification and evolution of prokaryotes and eukaryotes. In: King R.C. (eds) Bacteria, Bacteriophages, and Fungi, pp. 1-41. Springer, Boston, MA (1974) doi:10.1007/978-1-4899-1710-2_1
11. Yu, N., Yu, Z., Li, B., Gu, F., Pan, Y.: A comprehensive review of emerging computational methods for gene identification. JIPS 12(1), 1-34 (2016) doi:10.3745/JIPS.04.0023
12. Bandyopadhyay, S., Maulik, U., Roy, D.: gene identification: classical and computational intelligence approaches. IEEE TSMCC 38(1), 55-68 (2008) doi:10.1109/TSMCC.2007.906066
13. Claverie, J.: Computational methods for the identification of genes in vertebrate, genomic sequences. Human Molecular Genetics 6(10),1735-1744 (1997)
14. Picardi, E., Pesole, G.: Computational methods for ab initio and comparative gene finding. Methods in Molecular Biology 609, 269-284 (2010) doi:10.1007/978-1-60327-241-4_16
15. Math C., Sagot M., Schiex T., and Rouz P.Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research, 30(19), 41034117, (2002)
16. Gunawan, T.S, Epps, J., Ambikairajah, E.: Boosting approach to exon detection in DNA sequences. Electronics Letters 44(4), 323-324 (2008) doi:10.1049/el:20082343
17. Meyer, M., Durbin, R.: Comparative ab initio prediction of gene structures using pair HMMs. Bioinformatics 18(10), 13091318 (2002)
18. Burge, C., Karlin, S.: Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology 268(1), 78-94 (1997)
19. Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., Gnirke, A.: Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. PNAS 106(9), 32643269 (2009) doi:10.1073/pnas.0812841106
20. Zhang, C.T., Wang, J.: Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. Nucleic Acids Research 28(14), 28042814 (2002)
21. Wang, Z., Chen, Y., Li, Y.: A brief review of computational gene prediction methods. Genomics, Proteomics & Bioinformatics 2(4), 216221 (2004)
22. Birney, E., Durbin, R.: Using GeneWise in the Drosophila annotation experiment. Genome Research 10(4), 547548 (2000)
23. Mignone, F.: Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. Nucleic Acids Research 31(15), 46394645 (2003) doi:10.1093/nar/gkg483
24. Meisler, M.H.: Evolutionarily conserved noncoding DNA in the human genome: How much and what for? Genome Research 11(10), 1617-1618 (2000) doi:10.1101/gr.211401
25. Min, X.J., Butler, G., Storms, R., Sang, A.T.: OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Research 33, W677-680 (2005)

26. Kan, Z.,Rouchka, E.C., Gish, W.R., States, D.J.: Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Research 11(5), 889-900 (2001)
27. Gelfand, M.S.: Gene recognition via spliced sequence alignment. PNAS 93(17), 9061-9066 (1996) doi:10.1073/pnas.93.17.9061
28. Birney, E., Durbin, R.: Using GeneWise in the Drosophila annotation experiment. Genome Research 10(4), 547-548 (2000) doi:10.1101/gr.10.4.547
29. Guigó, R., Agarwal, P., Abril, J.F., Burset M., Fickett, J.W.: An assessment of gene prediction accuracy in large DNA sequences. Genome Research 10(10),1631-1642 (2000) doi:10.1101/gr.122800
30. Olson, R.S., Cava, W., Mustahsan, Z., Varik, A., Moore, J.H.: Data-driven advice for applying machine learning to bioinformatics problem. Pacific Symposium on Biocomputing 23, 192-203 (2018)
31. Cheng, J.: Machine learning algorithms for protein structure prediction. PhD dissertation, California State University at Long Beach (2006)
32. Yip, K.Y., Cheng C., Gerstein M.: Machine learning and genome annotation: a match meant to be? Genome Biology 14(5), 205 (2013) doi:10.1186/gb-2013-14-5-205
33. Sacar, D., Allmer, J.: Machine learning methods for microRNA gene prediction. Methods in Molecular Biology 1107, 177-187 (2014) doi:10.1007/978-1-62703-748-8_10
34. Le, D.H., Xuan, H.N., Kwon, Y.K.: A comparative study of classification-based machine learning methods for novel disease gene prediction. Knowledge and Systems Engineering. AISC, vol. 326, pp. 577-588. Springer, Cham (2015) doi:10.1007/978-3-319-11680-8_46
35. Schneider, H.W., Raiol, T., Brigido, M.M., Walter, M.E.M., Stadler, P.F.: A support vector machine based method to distinguish long non-coding RNAs from protein coding transcripts. BMC Genomics 18(1), 804 (2017) doi:10.1186/s12864-017-4178-4
36. Song, Y., Liu C., Wang, Z.: A machine learning approach for accurate annotation of noncoding RNAs. IEEE/ACM TCBB 12(3), 551-559 (2015) doi:10.1109/TCBB.2014.2366758
37. Choudhary, R., Gianey, H.K.: Comprehensive review on supervised machine learning algorithms. In: MLDS 2017, pp. 37-43 (2017) doi:10.1109/MLDS.2017.11
38. Bauckhage, C., Drachen A., Sifa, R.: Clustering game behavior data. IEEE TCIAIG 7(3), 266-278 (2015) doi:10.1109/TCIAIG.2014.2376982
39. Dai, W., Xue, G., Yang, Q., Yu, Y.: Transferring naive Bayes Classifiers for Text Classification. In: AAAI 2007, pp. 540-545 (2007)
40. Liu, B., Blasch, E.,Chen, Y., Shen, D., Chen, G.: scalable sentiment classification for big data analysis using naïve bayes classifier. In: IEEE BigData 2013, pp. 99-104 (2013) doi:10.1109/BigData.2013.6691740

A Study on the Innovative Health Data Management System based on Distributed Ledger

Junho Moon[1] and Dongsoo Kim[1,*],

[1] Department of Industrial and Information Systems Engineering, Soongsil University
369 Sangdoro, Dongjak-Gu, Seoul, Korea
jhmoon@soongsil.ac.kr, dskim@ssu.ac.kr, *Corresponding author

Abstract. In this paper, we propose an innovative health data management method which can manage and store the own health data directly by the information subject. Personal Health Record (PHR) systems have been recognized as an effective tool for individuals to manage their own health. However, the PHR systems have a variety of technical problems. In addition, current operation methods of the PHR systems can cause legal problems in some countries. As a result, most of the PHR systems are not used as much as expected. There have been several studies that applied blockchain technology in PHR systems to solve these problems. However, due to the limitations of the blockchain technology, not all problems have been solved yet. Therefore, we propose a new health data management system that allows information subjects to store and manage their own health data on their own devices. It solves legal problems about privacy by storing and managing health data in user device. And the proposed system uses a new network structure and a central management server to improve network reliability and availability. The proposed system enables health data to be managed and utilized more securely and effectively.

Keywords: Personal Health Record, Health Data, R3 Corda, Distributed Ledger

1 Introduction

Health is a very important issue for us and we are working hard to keep it. As a result, medical technology has evolved steadily. And in recent years, the integration of ICT technology accelerated its development. The development of technology has led to a paradigm shift in the healthcare industry. In the past, the medical industry has focused on treatment and diagnosis, and in recent years it has turned into accurate and predictable medicine. With the start of the fourth industrial revolution, the healthcare industry aims to transform into a smart medical industry through the integration of new technologies such as IoT, AI, mobile, Big Data and the cloud. The main goal of the smart healthcare industry is 4P (Predictive, Preventative, Participatory, Personalized) [1,2,3].

Smart Healthcare's ultimate goal is to provide personalized healthcare services that require the integration of diverse data, such as individual medical records, life logs, and genomic information. As a result, personalized health information management methods became necessary, and the Personal Health Record (PHR) systems were a technology that has been performing this function for a long time [4,5]. The effectiveness and necessity of PHR systems for the development of the medical industry has been highly evaluated in many studies, and several institutions and companies developed the PHR systems. However, most of the PHR systems are not used by user as much as experts expect. As a result, providers of PHR systems are trying to make new changes with the integration of blockchain technology. Nevertheless, due to the limitations of blockchain technology, not all problems have been solved, and defects in the blockchain technology can cause new problems in the PHR systems. Therefore, in this paper, we propose a new health data management system based on distributed ledger. This system allows health data subjects and medical institutions store and manage data together. In addition, this system promotes user participation by distributing the benefits of using health data to users. As a result, personal health is improved and the healthcare industry is developed through the utilization of information.

2 Related Work

2.1 Personal Health Record (PHR) system

The Personal Health Record (PHR) is a term used since 1978, and definitions of PHRs are defined differently by various institutions. The American Health Information Management Association (AHIMA) defined the PHR as "The personal health record (PHR) is an electronic, universally available, lifelong resource of health information needed by individuals to make health decisions. Individuals own and manage the information in the PHR, which comes from healthcare providers and the individual. The PHR is maintained in a secure and private environment, with the individual determining rights of access. The PHR is separate from and does not replace the legal record of any provider [4]."

PHR systems are difficult to implement and require a lot of infrastructure. Therefore, the PHR systems was developed by big-brother of IT industry, such as Google, Apple, and Microsoft, and it was built into a platform-type lifelong health management infrastructure. However, the PHR systems using a centralized database has some problems, such as the difficulty in collecting scattered health information, standard mismatch between groups, risk of data forging and modulation, risk of data security, and lack of data utilization [6,7].

2.2 Blockchain in PHR systems

In November 2008, Satoshi Nakamoto proposed a way to build trust in transactions through proof-of-work for the Bitcoin paper. It solves the trust problem of distributed ledger technology [8]. PHR systems are trying to adopt blockchain technology to solve its problems. A variety of blockchain based PHR systems are being under development or in service such as MedRec, GemHealth, Health Bank, and MediBloc

[9]. Blockchain is used in two types in the information system. First is On-Chain method. Second is the Off-Chain method.

Table 1. Features of the blockchain method

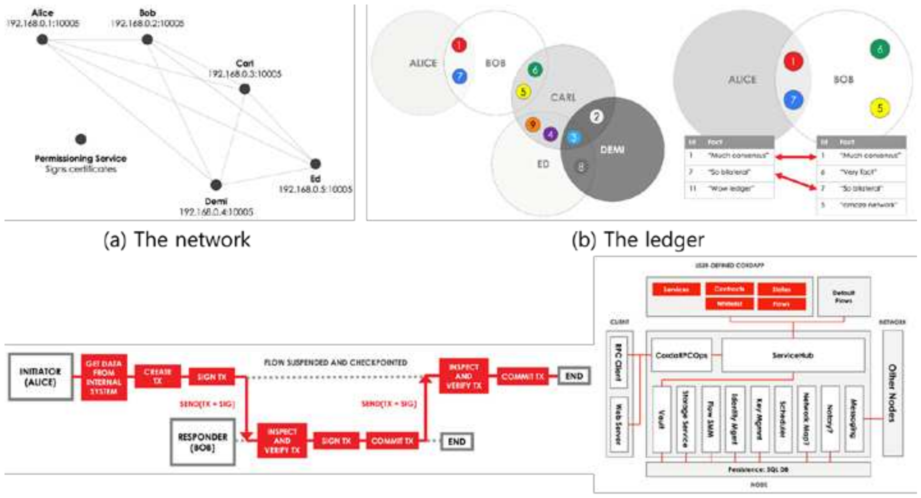
	On-Chain	***Off-Chain***
Data types	· Standardized data fields containing summary information in text form (e.g. age, gender)	· Expansive medical details (e.g. notes) and abstract data types (e.g. MRI images, human genome)
Pros	· Data is immediately visible and ingestible to all connected organizations, making blockchain the single source of truth	· Storage of any format and size of data
Cons	· Constrained in the type and size of data that can be stored	· Data is not immediately visible or ingestible, requiring access to each health care organization's source system for each record · Requires Off-Chain micro-services and additional integration layers · Potential for information decay on the blockchain

The difference between the two methods is shown in Table 1. The Off-Chain method have more disadvantages than the On-Chain. However, most of the blockchain based PHR systems use the Off-Chain type. This is because the shape and size of medical data are so diverse. And it is dangerous to store medical data outside the medical institution [6,10].

The adoption of blockchain technology was expected to solve the problems of traditional PHR systems. However, blockchain based PHR systems has problems that have not yet been solved. And, it has new problems due to the adoption of blockchain. First, the use of off-chain method can detect forgery and modulation of data. However, it cannot restore the original value. It requires a backup server to restore the data. Second, security problems of the blockchain are constantly being seen. Blockchain security issues can cause a large amount of data leakage, because the contents of the blockchain are shared by all participants in the network. Third, block mining costs are excessive. The proof of work of the blockchain requires a large amount of computer equipment and electric power. Adoption of blockchain can reduce the load on the central system. However, it can increase the overall system cost. Fourth, personal health data is not wholly owned by the data subject. The data of blockchain system is not stored on my device. The data is stored on an existing central system or on an unknown node in the network. Fifth, blockchain based PHR systems has the risk of losing user data due to loss of private key. The management of the private key is very important for the security of the blockchain system. When a user loses a key, he cannot find all his data in the blockchain network. Finally, storing health information in a blockchain network can be a legal issue in many countries. Health data is very sensitive data. Therefore, many countries have strict legal control. In the case of the Republic of Korea, health data cannot be kept outside the owner, medical institutions, and authorized government organization.

2.3 R3 Corda

R3 Corda is the primary reference technology for the proposed system design. R3 Corda is a type of consortium blockchain. It is very different in structure from other existing blockchain systems. Fig. 1 is the key concepts of Corda. Corda network is based on the P2P structure, as in Fig. 1-(a), and it have a node that provides permission services. Corda is different from other blockchain systems. It does not share data with all participants in the network. Corda shares data only with nodes participating in data generation as shown in Fig. 1-(b). Data consensus process of corda is different from other blockchain systems. As shown in Fig. 1-(c), Corda proves consensus of the data through agreements between trading partners. In addition, Corda verifies the transaction through the Notary function of the node architecture, as shown in Fig. 1-(d). Corda does not validate data in all nodes. It is done by using the user specified node or any node as the notary node [11].



(a) The network

(b) The ledger

(c) Example of flows

(d) Nodes

Fig. 1. Key concepts of R3 Corda

R3 Corda is well suited for the proposed system in this study in that not all nodes share data, only the transaction participants share data. In addition, unlike other blockchains, the consensus process is simplified and the burden on the system is minimized.

However, the Corda system has limitations in achieving all the goals of the proposed system. First, Corda does not provide data search function between network participants. Second, if the node is corrupted or deleted, the node cannot be recovered. Third, it is not suitable for use in a personal mobile device, such as the need for a static IP address and the installation of a server program.

Therefore, we have designed a new framework to achieve the goal of the proposed system. It is based on the structure of the R3 Corda system, eliminating unnecessary features on the nodes and designing additional functions needed to achieve the goal.

3 Architecture of health data management system

3.1 Goals of framework

The goals of the proposed framework are as follows: First, it ensures the reliability and integrity of the data. Second, data subjects can directly store health data. It is designed to store data on personal mobile devices for this purpose. Third, this system prevents data forgery and modulation and, it supports recovery of lost data. Fourth, this system creates a network between the data subject and the health data control institute. Fifth, this system is possible to search data between network members to improve health information utilization. Sixth, this system enables secure data transmission among network members.

3.2 Design of the framework

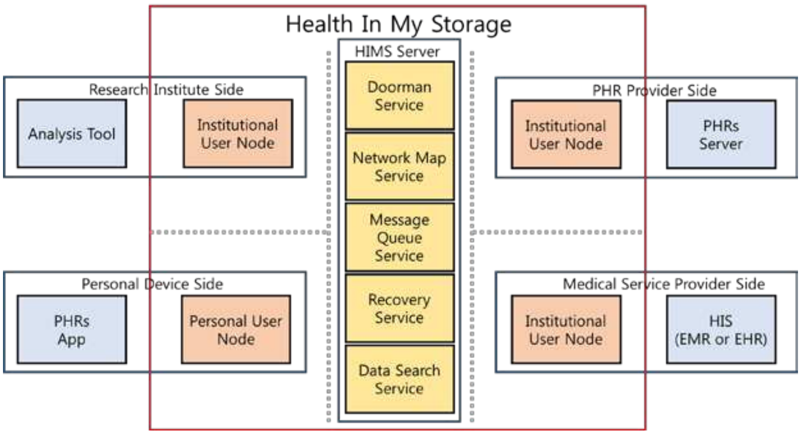


Fig. 2. Framework Overview

Fig. 2 is an overview of the proposed system. We named this system “Health in My Storage: HIMS.” As shown in Fig. 2, the HIMS is composed of three types of participants. First, the HIMS Server provides functions such as authentication of network participants, access path provisioning, message relaying, node restoration, data inquiry in network. Second, institutional user nodes include medical institutions, research institutions, and PHR providers. They record and use health data as a health data processor. In addition, the institutional user node acts as a notary to verify data on the network. Notary function ensures data reliability and integrity for the first goal of the framework. Third, a personal user node collects and manages health data. HIMS is designed to run on the user's mobile device. This is the second goal of the framework. Personal user nodes share their health data. It allows the medical industry to use the data. And, personal user node receives a portion of the profits generated by the data utilization. Personal user node also manages the authority for own health data. Health data is stored at a medical institution node or a PHR institution node and a personal user node. Health data is stored in two places. Therefore, faulty nodes can be recovered using data from another node. This is the third goal of the framework.

The proposed system distributes health data and verification data to related nodes. This enables functions such as securing data reliability, restoring faulty nodes, and sharing data. The system does not share data with all nodes like other blockchain systems. It shares data only with nodes related to data. It improves security.

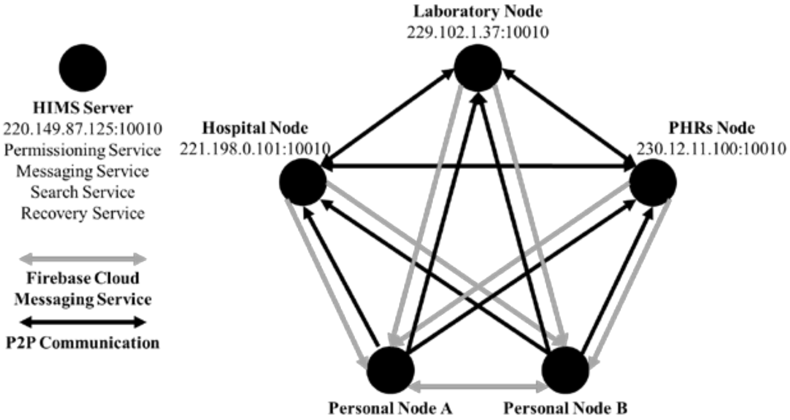


Fig. 3. Network Structure

The network of HIMS communicates as shown in Fig. 3. Every node of other blockchain systems requires a static IP address. However, HIMS uses personal mobile devices as personal user nodes by using Firebase Cloud Messaging (FCM). A HIMS server or an Institutional User Node requires a static IP address and a Web Application Server (WAS). This enables direct communication between nodes. On the other hand, FCM services should be used to access personal user nodes that do not have a static IP address and a WAS. The configuration of the new network allows a personal user node operating in the personal mobile device to become a member of the network. This configuration allows the framework's fourth goal to be achieved. In addition, the Firebase Cloud Messaging service enables broadcast messaging to all personal user nodes of the network. It can query data from all personal user nodes in the network easily and quickly. This function of the FCM allows to achieve the proposed fifth goal. All communications over the network are encrypted using RSA method. This can achieve the sixth proposal goal.

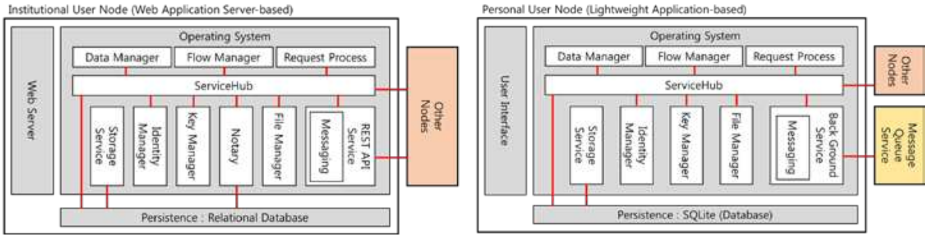
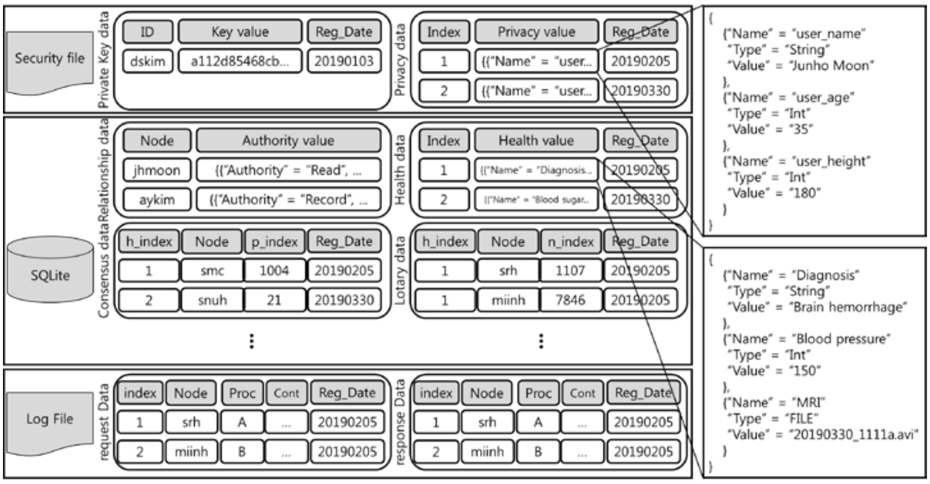


Fig. 4. Nodes Structure

The left side of Fig. 4 shows the configuration of the institutional user node. It was designed with reference to the Corda node structure. Each function of the node operates separately. This is to protect the node from external threats. All external requests to access the node are only possible with the REST API service. All processes are handled through ServiceHub for security, such as authentication and log records. The right side of Fig. 4 shows the configuration of a personal user node. Personal user nodes are also designed with reference to the structure of the Corda node. The personal user node includes only minimal functions for operate on the personal mobile device. And, it has a Back Ground Service to access the node from the outside.



ID	Key value	Reg_Date
dskim	a112d85468cb...	20190103

Index	Privacy value	Reg_Date
1	{{"Name" = "user...	20190205
2	{{"Name" = "user...	20190330

Node	Authority value
jhmoon	{{"Authority" = "Read", ...
aykim	{{"Authority" = "Record", ...

Index	Health value	Reg_Date
1	{{"Name" = "Diagnosis...	20190205
2	{{"Name" = "Blood sugar...	20190330

h_index	Node	p_index	Reg_Date
1	smc	1004	20190205
2	snuh	21	20190330

h_index	Node	n_index	Reg_Date
1	srh	1107	20190205
1	miinh	7846	20190205

index	Node	Proc	Cont	Reg_Date
1	srh	A	...	20190205
2	miinh	B	...	20190205

index	Node	Proc	Cont	Reg_Date
1	srh	A	...	20190205
2	miinh	B	...	20190205

```
{
  {"Name" = "user_name"
   "Type" = "String"
   "Value" = "Junho Moon"
  },
  {"Name" = "user_age"
   "Type" = "Int"
   "Value" = "35"
  },
  {"Name" = "user_height"
   "Type" = "Int"
   "Value" = "180"
  }
}
```

```
{
  {"Name" = "Diagnosis"
   "Type" = "String"
   "Value" = "Brain hemorrhage"
  },
  {"Name" = "Blood pressure"
   "Type" = "Int"
   "Value" = "150"
  },
  {"Name" = "MRI"
   "Type" = "FILE"
   "Value" = "20190330_1111a.avi"
  }
}
```

Fig. 5. Data Structure

Fig. 5 shows the data structure of a personal user node. The personal user node data is stored in three places. First, the security file stores the user's private key and privacy data. This file is managed by a separate security system. It can be controlled by only the device's owner. Second, SQLite is a relational database commonly used on mobile devices. This includes operational data such as authority information of another node designated by the user, health data, consensus data, verification data, and the like. The shape and size of health data is very diverse. It stores health data in JSON format to store flexible data. Storing data in JSON format does not take advantage of relational databases. To solve this problem, we are looking for ways to use relational database like a full-text search engine. Third, the log file records all activity on the node. This file is used to track problems of the node.

4 Conclusions

In this study, we looked at the traditional PHR systems for health data management and the adoption of blockchain technology to improve the PHR systems. And we have proposed an improved personal health data management system. In the proposed system, the personal users can participate directly in the network. Because it uses lightweight nodes for personal user node. And, personal users share the monetary benefits of the system as network participants. As a result, personal users actively participate in the system. It has also been proposed in existing blockchain based PHR systems. However, other proposed blockchain PHR systems may have legal problems in many countries. This proposed system can resolve legal disputes in most countries. This is because the data is stored and managed directly by the information owner. This system improves the utilization of health information and, as a result, promotes the personal health and medical industry.

Personal user nodes can cause security concerns due to security issues on the mobile device itself. However, mobile devices continue to evolve and this issue will be resolved soon. The proposed system does not yet have a health data standardization function. Therefore, we will further study how to standardize health data.

Acknowledgments. This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1D1A1B05029080).

References

1. Liao, W., Tsai. F.: Personalized medicine: A paradigm shift in healthcare. BioMedicine. vol. 3:2, pp. 66--72 (2013)
2. Golubnitschaja. O., Baban. B., Boniolo. G., Wang. W., Bubnov. R., Kapalla. M., Krapfenbauer. K., Mozaffari. M. S., Costigliola. V.: Medicine in the early twenty-first century: paradigm and anticipation - EPMA position paper 2016. EPMA Journal. vol. 7:23. (2016)
3. Hood. L., Auffray. C.: Participatory medicine: a driving force for revolutionizing healthcare. Genome Medicine. vol. 5:110. (2013)
4. AHIMA e-HIM Personal Health Record Work Group: The Role of the Personal Health Record in the HER. Journal of AHIMA. 76 (7). (2008)
5. Tang. PC., Ash. JS., Bates. DW., Overhage. JM., Sands. DZ.: Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption. J Am Med Inform Assoc. vol. 13, pp. 121—126. (2006)
6. Krawiec. RJ., Housman. D., White. M., Filipova. M., Quarre. F., Barr. D., Nesbitt. A., Fedosova. K., Killmeyer. J., Israel. A., Tsai. L.: Blockchain: Opportunities for Health Care. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/public-sector/us-blockchain-opportunities-for-health-care.pdf. (2016)
7. Moon. J., Lee. S., Byun. J., Choi. J. S., Kim. D.: A Case Study of Interconnected PHR System Implementation. ICIC Express Letters. vol. 10:6. pp. 501—508. (2019)
8. Nakamoto. S.: Bitcoin: A Peer-to-Peer Electronic Cash System. (2008)
9. Jeon. J., Kim. Y.: Case Study of Medical Record Management Platform using Block Chain. In: Korea Software Congress 2019. (2018)
10. Linn. LA., Koo. MB.: Blockchain for health data and its potential use in health IT and health care related research. In: ONC/NIST Use of Blockchain for Healthcare and Research Workshop. Gaithersburg, Maryland, United States, ONC/NIST. (2016)
11. R3: R3 Corda Development Documentation Version 4. https://docs.corda.net/releases/release-V4.0/. (2019)

ADeepTool: Application Tool for Alzheimer's Disease Diagnosis Based on Deep Learning Approach

Ngoc-Huynh Ho[1], Hyung-Jeong Yang[1,*], Ho-Chun Song[2], Jahae Kim[2]
[1] School of Electrical and Computer Engineering, Chonnam National University
61186, Gwangju, South Korea
[2] Department of Nuclear Medicine, Chonnam National University Hospital and Medical
School, Gwangju, South Korea
[*]Corresponding Author: hjyang@jnu.ac.kr

Abstract. In this study, we present a novel application tool for classification of Alzheimer's disease (AD) using deep learning model, called ADeepTool. The tool includes two properties: preprocessing and diagnosis. The preprocessing property contains four steps to generate desired image for AD prediction. The first step is co-registration that refers to the spatial alignment of a series of the positron emission tomography (PET) images from dimension of the magnetic resonance (MR) images. The second step calls segmentation, that extracts the binary masks of the grey matter (GM) and white matter (WM) tissues from MRI images. Then, the inversion of GM / WM mask is mapped into the co-registered PET images in the third step. Finally, the non-GM / non-WM PET image is normalized in the last step to determine regional correspondence from the brain. For diagnosis property, the normalized PET brain is fed into the VGG16 model, which uses transfer learning of pretrained parameters from the high-level layer concatenation autoencoder (HiLCAE) model, to classify AD, mild cognitive impairment (MCI), and normal control (NC) subjects. The predicting model is trained with the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Also, we provide three-dimensional (3D) and four-dimensional (4D) visualization.

Keywords: Alzheimer's díease, deep learning, VGG16, HiLCAE.

1 Introduction

Nowadays, medical images have become essential in medical diagnosis and treatment. These images play an important role in medical applications because doctors are interested in exploring the internal anatomy [1]. Many techniques have been developed based on X-ray and cross-sectional images such as computed tomography (CT), magnetic resonance (MR) Imaging, single photon emission tomography (SPECT), positron emission tomography (PET), or ultrasound) [2, 3].

In medical research, Alzheimer's disease (AD) is a neurological disorder that causes memory loss and dementia. It is mainly observed in elderly individuals over the age of 60 but can also be caused by concussions or traumatic brain injuries. It causes brain cells to die and spread the damage across the brain, in some severe cases rendering an individual unable to perform daily necessary tasks. It is also considered as neurodegenerative type of dementia. To diagnose AD, doctors evaluate patient's signs and symptoms and conduct several tests.

According to these facts, in this study, we proposed an accurate and robust tool to automatically predict AD from medical images, called ADeepTool. The ADeepTool includes two properties: preprocessing and diagnosis. The preprocessing property contains four steps to generate desired image for AD prediction. In the first step, we co-register the space of a series of PET images from dimension of MR images. The second step calls segmentation, that we extract the binary masks of the grey matter (GM) and white matter (WM) tissues from MR images. Then, we invert the GM / WM mask to map into the co-registered PET images in the third step. Finally, the non-GM/non-WM PET images is normalized to determine regional correspondence from the brain. These steps are implemented by the statistical parametric mapping (SPM) toolbox version 12 [4]. For diagnosis property, the normalized PET images is fed into the VGG16 model, which uses transfer learning of pretrained parameters from the high-level layer concatenation autoencoder (HiLCAE) model [5], to classify AD, mild cognitive impairment (MCI), and normal control (NC) subjects.

The rest of this paper is organized as follows. Section 2 describes the reference methodology and comparison of conventional methods. Demostration of ADeepTool is shown in Section 3. Finally, Section 4 includes the conclusion and our future works.

2 Methodology and Comparison

In this section, we introduce methodology to develop the AdeepTool and conparison between conventional methods.

The proposed tool is implemented based on the method proposed by our previous study [5]. To summary, we introduce briefly the methodology. The dataset that we use to train model is downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [6].

First, we perform the four steps of the preprocessing, shown in Figure 1, as follows:

1. Co-registration - Co-register PET images to space of MRI images to result in same space and orientation for PET and MRI.
2. Segmentation - Segment MRI to WM or GM probability maps using an a brain template from the Monteal Neurological Institute (MNI).
3. Mapping - Extract WM or GM from PET image by the order of sub-steps:

   3.1 Inversely binarize WM-map or GM-map of MRI using thresholding. To choose a suitable threshold, we plot normalized histogram of all WM/GM images and determine the average value where the dark (background) and bright (WM/GM) regions are separated.

   3.2 Cover the full PET image by the mask in step 3.1. It results in a "non-WM/GM PET" image
4. Normalization - Normalize "non-WM PET" or "non-GM PET" images to a standard MNI template, using transformation matrix that calculated from normalizing coupled MRI to an MRI template in MNI space.

The MNI template is a template that refers to a representative image with anatomical features in a coordinate space, which then provides a target to individual images aligned to. After preprocessing, the normalized images are divided into several patches for pretraining layers using the High-level Layer Concatenation Autoencoder (HiLCAE) architecture [5], as shown in Figure 2. This network is a variant of convolutional autoencoder (CAE) network. The differences between the HiLCAE and traditional CAE is that there is a concatenation on the high resolution features from the encoding layer with the corresponding decoding layer. A successive convolution layer can then learn to assemble a more precise output based on this information. Moreover, the deconvolution part has also a large number of feature channels, which allow the network to propagate context information to higher resolution layers. Then, the 3D-VGG16 model in Figure 3 is used to identify the three states of AD diagnosis. The first four convolutional layers are set up by the learned hyper-parameters (weights and bias) from pre-trained HiLCAE, while the rest of layers use random initialization. The results on ADNI dataset were presented in [5]. Table 1 summarize the outstanding results of AD detection.

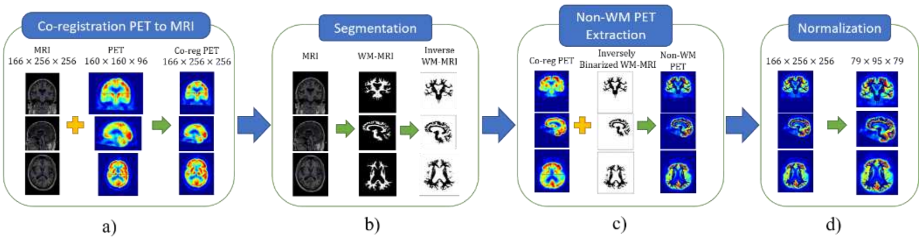


Fig. 1 Preprocessing step. a) Co-registration of PET dimention to MRI space; b) Segmentation of GM/WM from MRI and inversion; c) Exclusion of GM/WM from PET; d) Normalization of excluded PET to the same dimention

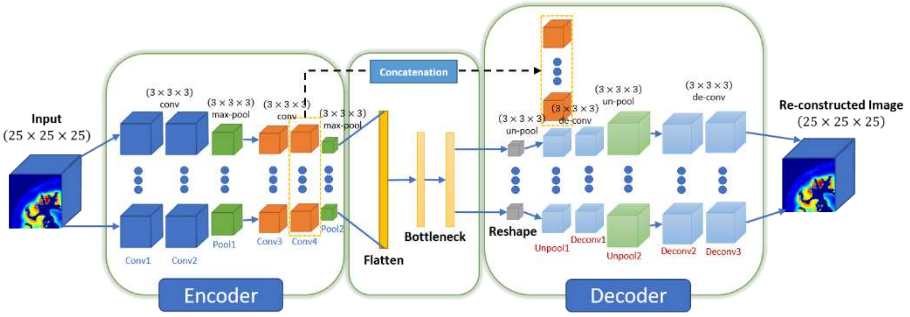


Fig. 2 Pre-training HiLCAE model

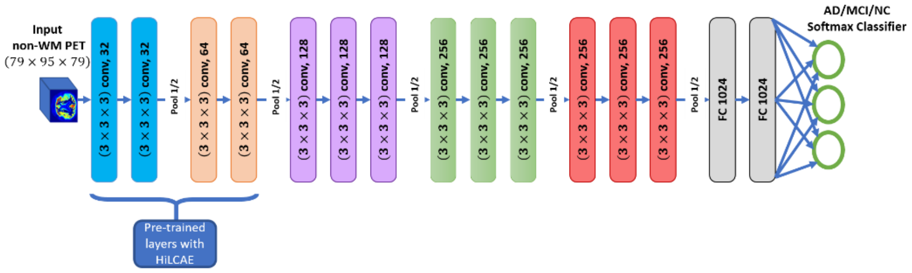


Fig. 3 3D-VGG16 model for AD classification

Table 1. Comparison results for AD classification between previous methods

Method	AD vs MCI	MCI vs NC	AD vs NC	AD vs MCI vs NC
Gupta et al. [7]	88.1	86.35	94.75	85
Suk et al. [8]	83.7	90.7	98.8	-
Vu et al. [5]	**93**	**95**	**98.8**	**91.13**

3 Materials and Applications

This ADeepTool was completed based on the MATLAB App Designer (2018b) for developing interface and functions. AD prediction was implemented on PYTHON framework. Fig. 4 shows the general diagram of ADeepTool. There are two main interfaces. The first one is “Preprocessing” which includes the four steps based on SPM function to extract the fused modality between MR and PET images.

Fig. 5 presents the interface of preprocessing tab, which includes a button for loading patient information file, a list box of patient IDs, a group of radio buttons and spinners for selecting axes view and slides, an adding new patient button, and four panels of 3D visualization corresponding to four steps. Inside each panel, the “process” button is used to execute the role of the corresponding step.

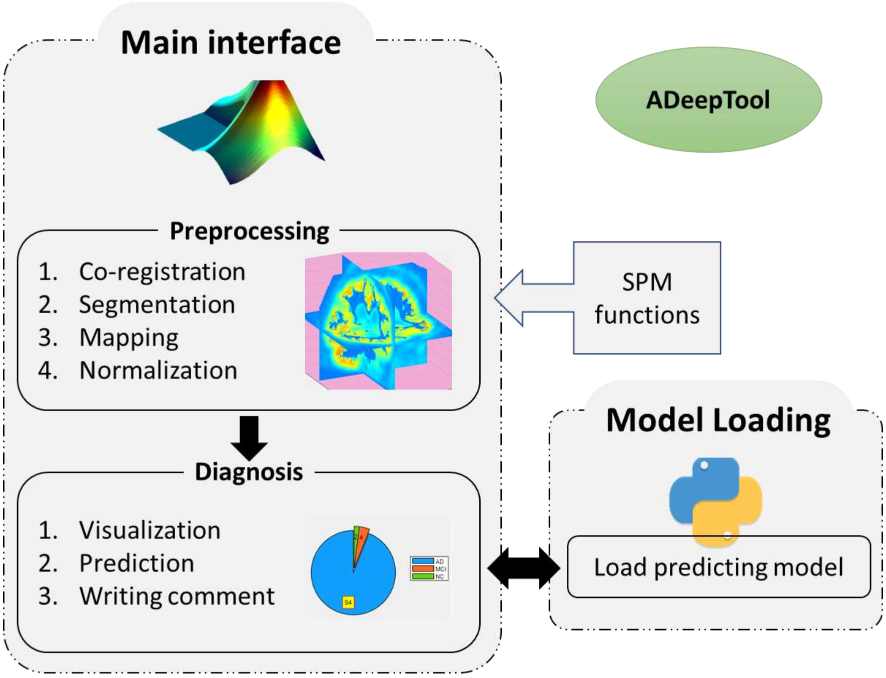
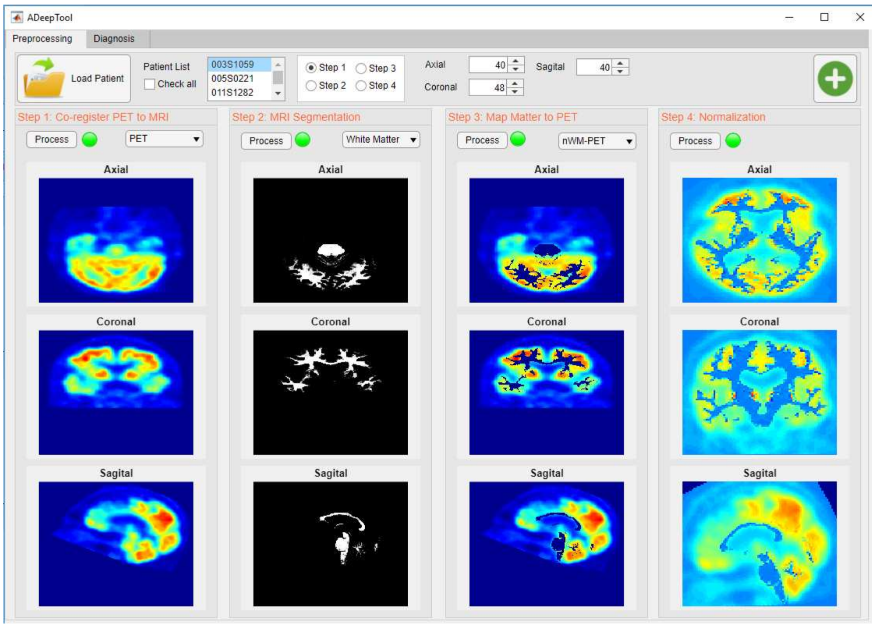


Fig. 4 ADeepTool diagram for AD detection



Patient List: 003S1059, 005S0221, 011S1282

□ Check all

● Step 1 ○ Step 2 ○ Step 3 ○ Step 4

Axial 40 Coronal 48 Sagital 40

Step 1: Co-register PET to MRI — PET

Step 2: MRI Segmentation — White Matter

Step 3: Map Matter to PET — nWM-PET

Step 4: Normalization

Fig. 5 Preprocessing interface.

Fig. 6- 1 to 6- 4 illustrate the interface of diagnosis tab, which includes a load patient information button, a load predicting model button, a view button, a save doctor's comment button, a group of three sliders for changing slides, a list box of patient IDs and searching ID box, four state button for choosing MRI or PET with GM or WM view, a region for image display, a region for result display, and a text box that doctor can write some description about disease condition. The non-GM and non-WM views are available if and only if the PET view is chosen.

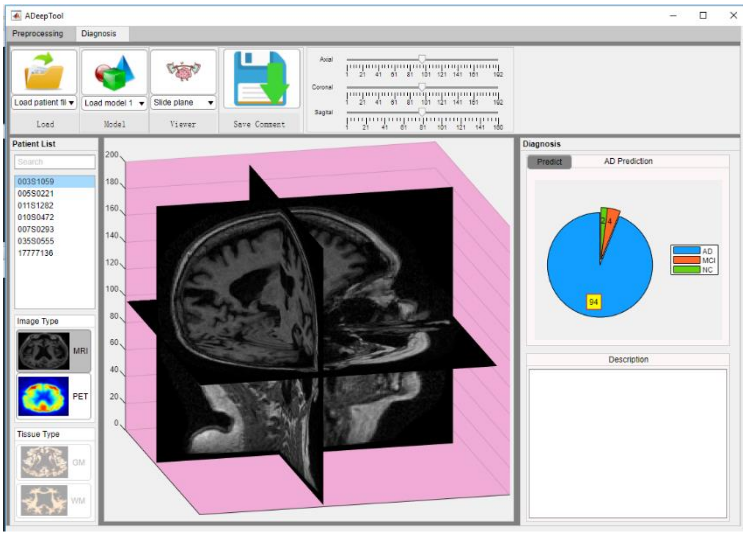


Fig. 6- 1 AD case with MRI view

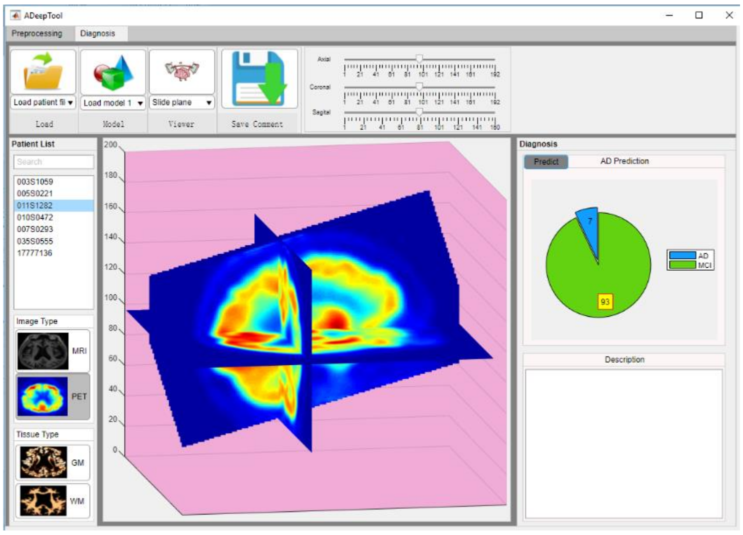


Fig. 6- 2 MCI case with PET view

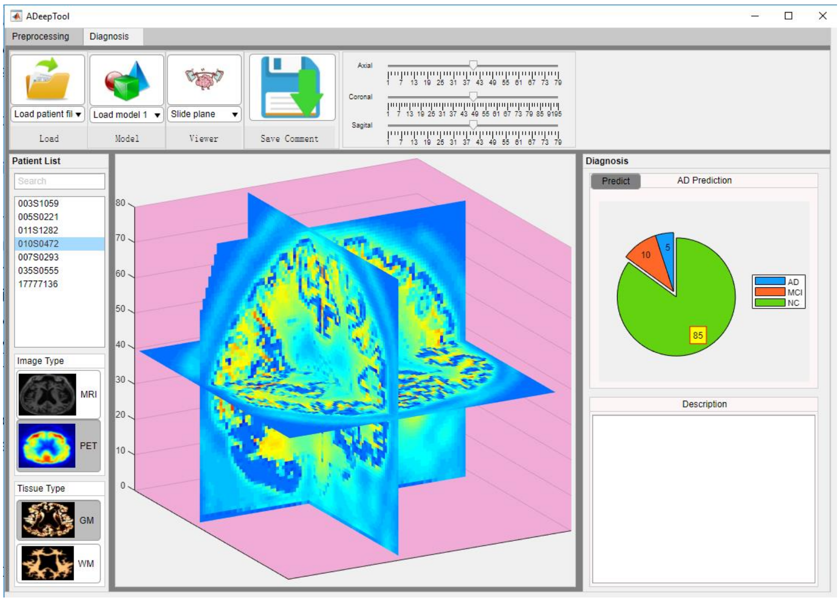


Fig. 6- 3 NC case with non-GM PET view

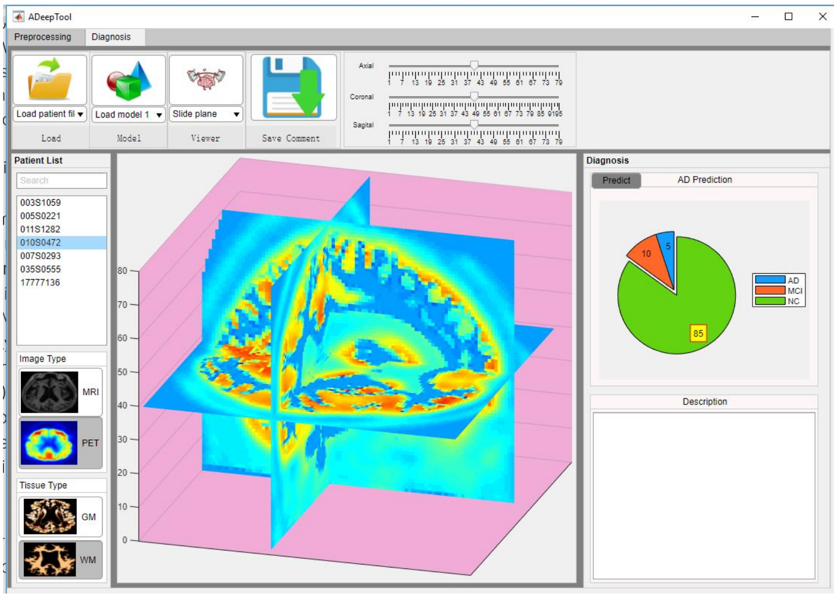


Fig. 6- 4 NC case with non-WM PET view

4 Conclusion and Future Works

This paper study presents a robust and automatic ADeepTool application for AD classification based on deep learning. There are two main features developed: preprocessing medical images and classifying AD. The 3D and 4D visualization are included for doctor's view. For future development, statistic computations (e.g. SUV) will be included to provide more information for doctor's medical diagnosis. Also, region of interest (ROI) or overlaying visualization will be progressed.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2017R1A2B4011409)

References

1. T. Saba, S. Alzorani, and A. Rehman, "Expert system for offline clinical guidance and treatment," Life Sci. J., Vol. 9, no. 4, pp. 2639-2658, 2012.
2. A. Norouzi, A. Rahman, M. Shafry, and et al., "Visualization and segmentation," Int. J. Acad. Res., Vol. 4, no. 2, pp. 202-208, 2012.
3. J. Maintz, and M. Viergever, "A survey of medical image registration," Med. Image Anal., Vol. 2, no. 1, pp. 1-36, 1998.
4. https://www.fil.ion.ucl.ac.uk/spm/
5. T-D. Vu, and N-H Ho, and et al., "Non-white matter tissue extraction and deep convolutional neural network for Alzheimer's disease detection," Soft Computing, Vol. 22, no. 20, 2018.
6. http://adni.loni.usc.edu/data-samples/access-data/
7. A. Gupta, and M. Ayhan, and A. Maida, "Natural image bases to represent neuroimaging data," The 30th international conference on machine learning, pp 987–994, 2013.
8. H-I. Suk, and S-W. Lee, and D. Shen, "Latent feature representation with stacked auto-encoder for ad/mci diagnosis," Brain Struct Funct, Vol. 220, pp 841–859, 2015.

Scheduling Optimization and Reduce Turnaround Time by Analyzing HPC Job Logs

JunWeon Yoon, TaeYoung Hong

Supercomputing Center, KISTI, Gwahak-ro, Yuseong-Gu, Daejeon,
Republic of Korea,
{jwyoon, tyhong}@kisti.re.kr

Abstract. Parallel computing plays the role to solve the large-scale arithmetical problems using many distributed computing resources. The point of parallelism is to reduce the total execution time to solve the huge job. Also, it can increase the scale and complexity of the problem.

In this paper, we analyze user job logs for a certain period and recognize that the whole execution time is delayed due to the large-scale job. Turnaround time can be reduced by optimizing the inefficiency of waiting resources. Therefore, we applied the back-filling algorithm and confirmed that the overall execution time was reduced by placing small jobs in the latter priority on the available resources without delaying the original job in the queue.

Keywords: HPC, Supercomputer, Scheduling, Parallel computing, Backfilling

1 Introduction

The batch job scheduler recognizes the computational resources configured in the cluster environment and plays a role in efficiently arranging the jobs in order. The basic function of the scheduler is to accurately reflect the resource status [1]. This includes the various computational resources such as CPU, and memory, as well as various architectures such as GPU and Intel PHI. It also meets requirements such as the fair-share policy to ensure fair distribution of resource distribution after awareness of resources, preemption support for high-priority jobs, resource scalability assurance, and support for various system environments. Most job scheduler software reflects the user job environment, from job submission to termination, as well as the state of the inventory and system status of the entire managed object. Tachyon2 is Linux based cluster system with 3,200 computing nodes. Each node is equipped with two-socket Intel Xeon X5570 2.93GHz (Nehalem) and 24GB DDR3 Memory [2]. This system uses Sun Grid Engine (SGE) as a batch job scheduler [3].

This study analyzes the job execution logs performed in the batch scheduler. This log contains user information, execution time, resource size, job exit status, and so on. Also, the results of the log analysis show that the inefficiencies of waiting resources due to large-scale operations. Therefore, it can be seen that optimization can be performed by applying the backfilling. In conclusion, job turnaround time can be reduced by optimizing the inefficiency of waiting resources.

2 Related Work

2.1 Job Execution Logs

Tachyon2 stores job information at the start and end point of the job. When a job is submitted, it is queued and allocated resources, and then it is started. At this time, the user's job script is backed up, and the execution environment of the library, the compiler, and the node information allocated thereto are stored. When the job is completed, the SGE log is extracted and reprocessed, and the execution history is stored in the format shown in Table 1. This leaves a log of job information such as job execution date, job ID, Job-Name, Submit-Date, Start-Date, End-Date, Wait-Time, Execution-Time, the number of used CPUs, Exit-Code, Failed-Code and the number of threads. In this table, the CPU and the execution time (wait, start and end) fields of the job were analyzed to extract large-scale job statistics.

Table 1. Executed job information log of SGE

1	2	3	4	5
DATE	JOBID	GID	UID	JOBNAME
20180821	271481	gid032	uid000	mpi_solver
6	7	8	9	10
QNAME	SUBMIT (DATE)	START (DATE)	END (DATE)	WAIT(s)
normal	20180821	20180821	20180821230	201536
11	12	13	14	15
RUN(s)	CPUS	CPU USAGE	MEM USAGE	MAXV MEM
31852	80	22281.74	28.41	1876.06
16	17	18	19	20
STATUS	E-CPU	E-RUN(s)	EXIT-CODE	FAILED
D	80	2868160	0(32)	0(32)
21	22			
OMP_NUM_THREADS	TASK_NUM			
1	0			

2.2 Analysis of Job Execution Logs

In this section, we conduct research to maximize resource utilization through analysis of large-scale jobs. First, we divided the number of monthly jobs performed from January 2016 to December 2017 for large-scale resources use of more than 2,000 cores. Figure 1 shows statistics of public queue available nodes.

- Statistic results of the resource usage (Jan, 2016 ~ Dec, 2017)
▪ Average of waiting time: 9.3 hours.
▪ Average of node usage rate: 85.69 %

Job running in Tachyon2 is simply a first come first serve (FCFS) method in which jobs submitted first to the queue are executed first. However backfilling algorithms that can efficiently use resource fragmentation have not yet been applied. The average waiting time of the public queue is 9.3 hours, but the overall node utilization is only about 85.69%. The reason is that when a large-scale job is submitted, it will free up the resources it requires before execution. A job requires a small amount of resources if the job priority is low, it must wait until the large job is finished [4]. FCFS is the best way to guarantee the fairness of the job order, but, as the size of computational resources increases, there is a limit to the efficient use of resources. The simulation in the following section gives more details.

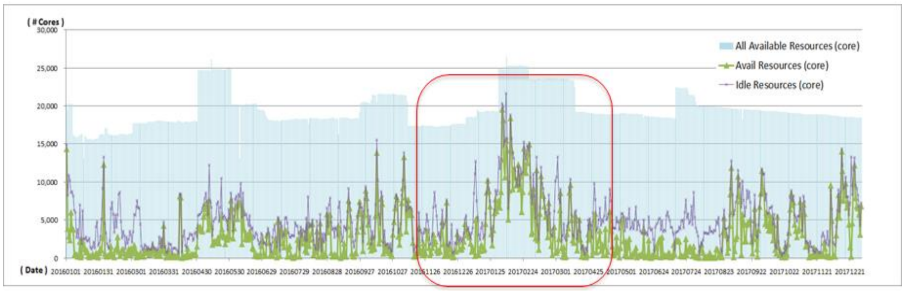


Figure 1. Statistics of public queue available nodes (Jan. 2016~Dec. 2017)

Figure 1 shows the trend of increasing available resources as the number of large-scale operations increases towards the end of 2016 and early 2017. This is because the resources required for each job are different and resources are allocated according to the arrival time of the job. Therefore, actual available resources are not used and fragmentation occurs.

The most extreme way to reduce resource fragmentation is Shortest Job First (SJF) algorithm, which is a way to prioritize small jobs according to fragmented resources. This can improve resource utilization and improve overall performance, but it does not guarantee fairness of job order and in the worst case can lead to starvation of large jobs. Therefore, the scheduling policy is used in a way that combines FCFS and SJF considering resource size and job characteristics [5].

3 Simulation

Backfill scheduling is a method of rearranging the order when a small job cannot be executed due to a relatively large predecessor job. This can improve performance while maintaining some fairness of the job. In backfill scheduling, the execution time of each job must be specified. In this paper, we show how the overall execution time, turn is reduced by applying a basic backfilling algorithm [6].

3.1 Backfilling Algorithm

The Conservative Backfilling algorithm is an early version, and it adheres to the FCFS scheme, which is the basic principle of scheduling, and running first when a subordinate job satisfies a fragmented resource. In order to implement this algorithm, a queued job must have the resource and time requirements. For this mechanism, there are two data structures. The first is a list structure that stores jobs and execution time of the queue list and the second is the resource processor profile to be used.

This algorithm requires no latency for prior jobs due to subordinate jobs, but cannot guarantee the planned sequence of jobs if the preceding job is terminated earlier than the expected time. Table 2 shows the Conservative Backfilling algorithm.

Table 2. Conservative backfilling algorithm

1) Search for the start time
   - ① Find the required resources and find the first point of available processors that have enough work to do.
   - ② Start from this point and continue to find out whether the processor is available as expected until the end of the operation.
   - ③ If not, go back to ① and continue scanning the next possible anchor point.

2) Updating the information of the waiting queue resource details reflecting the allocated details of the processors from the start point to the end point of the job

3) If the point of operation is the current point, perform the job immediately

3.2 Resource Efficiency

We simulate a specific large-scale job executed in Tachyon2. This system applied the job priority policy, but backfill scheduling is not applied. This experiment shows that the conservative backfilling (vanilla version) is applied to T2 to reduce overall execution time. As referred, the queued job has two properties which are the resource and time requirements. Here, it is assumed that the priority is higher as the number is smaller. If job #0 is inserted into the queue list as shown in Figure 2, it should wait until it has a suitable resource. The conservative backfilling can execute low priority jobs (box #1~#6) first during the waiting time of job #0. However, job 0 should not delay the start time.

In this experiment, we selected a large job requiring 2,048 nodes (16,384 cores) and the job executed after approximately 40 hours and 16 minutes of waiting time. During the waiting time, the computing nodes are emptied to match the requested resources of the. As a result, the waiting time of the entire job is increased. If a small job with a low priority leapfrogs first for the resources that are emptied during the waiting time, it can use the resources efficiently. The simulation proceeds as shown in Figure 3.

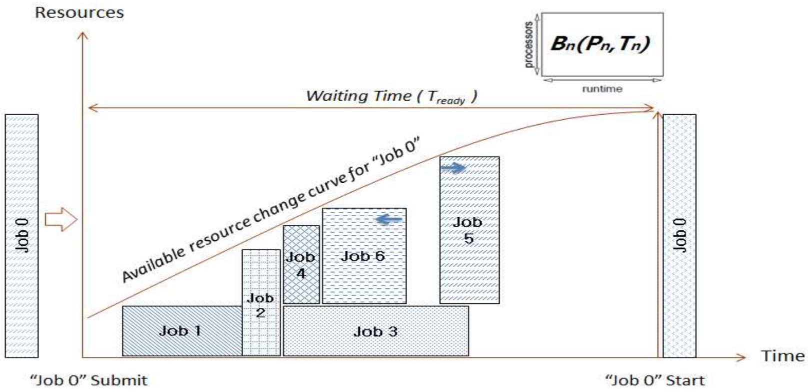
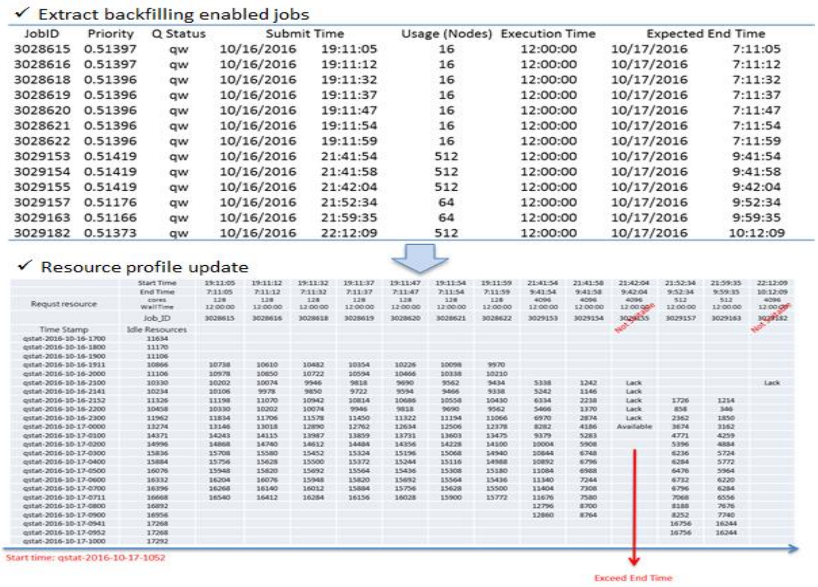


Figure 2. Apply backfilling to available resources before large work



✓ Extract backfilling enabled jobs

JobID	Priority	Q Status	Submit Time		Usage (Nodes)	Execution Time	Expected End Time	
3028615	0.51397	qw	10/16/2016	19:11:05	16	12:00:00	10/17/2016	7:11:05
3028616	0.51397	qw	10/16/2016	19:11:12	16	12:00:00	10/17/2016	7:11:12
3028618	0.51396	qw	10/16/2016	19:11:32	16	12:00:00	10/17/2016	7:11:32
3028619	0.51396	qw	10/16/2016	19:11:37	16	12:00:00	10/17/2016	7:11:37
3028620	0.51396	qw	10/16/2016	19:11:47	16	12:00:00	10/17/2016	7:11:47
3028621	0.51396	qw	10/16/2016	19:11:54	16	12:00:00	10/17/2016	7:11:54
3028622	0.51396	qw	10/16/2016	19:11:59	16	12:00:00	10/17/2016	7:11:59
3029153	0.51419	qw	10/16/2016	21:41:54	512	12:00:00	10/17/2016	9:41:54
3029154	0.51419	qw	10/16/2016	21:41:58	512	12:00:00	10/17/2016	9:41:58
3029155	0.51419	qw	10/16/2016	21:42:04	512	12:00:00	10/17/2016	9:42:04
3029157	0.51176	qw	10/16/2016	21:52:34	64	12:00:00	10/17/2016	9:52:34
3029163	0.51166	qw	10/16/2016	21:59:35	64	12:00:00	10/17/2016	9:59:35
3029182	0.51373	qw	10/16/2016	22:12:09	512	12:00:00	10/17/2016	10:12:09

✓ Resource profile update

	Start Time	19:11:05	19:11:12	19:11:32	19:11:37	19:11:47	19:11:54	19:11:59	21:41:54	21:41:58	21:42:04	21:52:34	21:59:35	22:12:09
	End Time	7:11:05	7:11:12	7:11:32	7:11:37	7:11:47	7:11:54	7:11:59	9:41:54	9:41:58	9:42:04	9:52:34	9:59:35	10:12:09
Requst resource	cores WallTime	128 12:00:00	128 12:00:00	128 12:00:00	128 12:00:00	128 12:00:00	128 12:00:00	128 12:00:00	4096 12:00:00	4096 12:00:00	4096 12:00:00	512 12:00:00	512 12:00:00	4096 12:00:00
	Job_ID	3028615	3028616	3028618	3028619	3028620	3028621	3028622	3029153	3029154	3029155	3029157	3029163	3029182
Time Stamp	Idle Resources													
qstat-2016-10-16-1700	11634													
qstat-2016-10-16-1800	11170													
qstat-2016-10-16-1900	11106													
qstat-2016-10-16-1911	10866	10738	10610	10482	10354	10226	10098	9970						
qstat-2016-10-16-2000	11106	10978	10850	10722	10594	10466	10338	10210						
qstat-2016-10-16-2100	10330	10202	10074	9946	9818	9690	9562	9434	5338	1242	Lack			Lack
qstat-2016-10-16-2141	10234	10106	9978	9850	9722	9594	9466	9338	5242	1146	Lack			
qstat-2016-10-16-2152	11326	11198	11070	10942	10814	10686	10558	10430	6334	2238	Lack	1726	1214	
qstat-2016-10-16-2200	10458	10330	10202	10074	9946	9818	9690	9562	5466	1370	Lack	858	346	
qstat-2016-10-16-2300	11962	11834	11706	11578	11450	11322	11194	11066	6970	2874	Lack	2362	1850	
qstat-2016-10-17-0000	13274	13146	13018	12890	12762	12634	12506	12378	8282	4186	Available	3674	3162	
qstat-2016-10-17-0100	14371	14243	14115	13987	13859	13731	13603	13475	9379	5283		4771	4259	
qstat-2016-10-17-0200	14996	14868	14740	14612	14484	14356	14228	14100	10004	5908		5396	4884	
qstat-2016-10-17-0300	15836	15708	15580	15452	15324	15196	15068	14940	10844	6748		6236	5724	
qstat-2016-10-17-0400	15884	15756	15628	15500	15372	15244	15116	14988	10892	6796		6284	5772	
qstat-2016-10-17-0500	16076	15948	15820	15692	15564	15436	15308	15180	11084	6988		6476	5964	
qstat-2016-10-17-0600	16332	16204	16076	15948	15820	15692	15564	15436	11340	7244		6732	6220	
qstat-2016-10-17-0700	16396	16268	16140	16012	15884	15756	15628	15500	11404	7308		6796	6284	
qstat-2016-10-17-0711	16668	16540	16412	16284	16156	16028	15900	15772	11676	7580		7068	6556	
qstat-2016-10-17-0800	16892								12796	8700		8188	7676	
qstat-2016-10-17-0900	16956								12860	8764		8252	7740	
qstat-2016-10-17-0941	17268											16756	16244	
qstat-2016-10-17-0952	17268											16756	16244	
qstat-2016-10-17-1000	17292													

Not Suitable

Not Suitable

Start time: qstat-2016-10-17-1052

Exceed End Time

Figure 3. Apply backfilling for a large scale job

Backfilling job have two properties Resource usage (P), Runtime (T). Therefore, the backfill job notation can be expressed as Eq.1. The resource efficiency by backfilling scheduling is measured as follows.

- Backfilling jobs: $B_1(P_1, T_1)$, ⋯, $B_n(P_n, T_n)$ (Eq.1)
- Resource Efficiency $= \sum_{i=1}^{n} P_i T_i$
- (128cores∗12hours)∗7jobs + (4,096cores∗12hours)∗2jobs + (512cores∗12hours)∗2jobs = 72,192seconds (Reduced overall execution time)

4 Conclusion and Future Work

Parallel computing is a suitable solution for solving large-scale problems [7]. The main content of this study is to optimize the order of jobs executed in the Tachyon2 system to efficiently use the entire available resources. This experiment analyzed the actual logs of job execution using the converted log in the Tachyon2 system. Note that the average utilization of the Tachyon2 system is about 85.6% and the average wait time is about 9.3 hours. This can be seen as fragmentation of resources that occurs during scheduling. Of course, it is not possible to use all the resources perfectly to reflect usage and execution time. However, it is possible to minimize resource fragmentation occurring during job scheduling and to make resource utilization more efficient by analyzing the jobs performed on the computational resources and studying and applying the appropriate scheduling algorithms. One way is to analyze the statistics of large-scale operations, which are gradually increasing in the Tachyon2 system, to grasp the user's success rate, execution time, and resource size. Based on the statistical results, the backfilling scheduling algorithm is studied and simulated to reduce resource fragmentation in Tachyon2.

In the future, we will apply the various scheduling algorithm to the current system and analyze the work execution history continuously and repeatedly. This can optimize resource utilization and reduce overall user latency. This research was conducted using a KISTI supercomputer. Currently, KISTI's 5th supercomputer, Nurion, is made up of 8,400 computing nodes based on the many-core architecture and launched production service in December 2018. The above study is proceeding in the same way for the 5th supercomputer and will be the basis for efficient resource utilization.

References

1. Abawajy JH.: An efficient adaptive scheduling policy for high-performance computing. Future generation computer systems, 25(3):364-370 (2009)
2. KIST National Supercomputing Center, http://www.ksc.re.kr/
3. Chaubal C: Scheduler Policies for Job Prioritization in the Sun N1 Grid Engine 6 System, Technical report, Sun Microsystems, Inc., Santa Clara, CA, USA (2005)
4. Iqbal S, Gupta R, Fang YC.: Planning considerations for job scheduling in HPC clusters. Dell Power Solutions:133-136 (2005)
5. Yuan, Y, Wu Y, Zheng W, Li K.: Guarantee strict fairness and utilize prediction better in parallel job scheduling. IEEE Transactions on Parallel and Distributed Systems, 25(4):971-981 (2014)
6. Feitelson DG, Weil AMA.: Utilization and predictability in scheduling the IBM SP2 with backfilling. In Parallel Processing Symposium and Proceedings of the First Merged International and Symposium on Parallel and Distributed Processing, IEEE: 542-546 (1998)
7. Yoon, J., Hong, T., Choi, J., Park, C., Kim, K., & Yu, H.: Evaluation of P2P and cloud computing as platform for exhaustive key search on block ciphers. Peer-to-Peer Networking and Applications, 11(6):1206-1216 (2018)

A method to provide homepage-free computing cluster monitoring using Google Sheets

Geonmo Ryu[1], Byungyun Kong[1], Heejun Yoon[1*] and Seo-Young Noh[2]

[1] Korea Institute of Science and Technology Information (KISTI), 245, Daehak-ro,
Yuseong-gu, Daejeon, Republic of Korea
{geonmo, kong91, k2}@kisti.re.kr
[2] Chungbuk National University, Chungdae-ro 1, Seowon-Gu, Cheongju, Chungbuk 28644,
Republic of Korea
rsyoung@cbnu.ac.kr

Abstract. In Korea, it is very inefficient to create a homepage that provides simple information by administrative and public agency website operating guidelines. So, we suggest using Google Sheets to replace web pages. Users can quickly reprocess the collected information on Google Sheets. Therefore, it is beneficial for users who have not been trained to develop computer programming. Besides, after a government research program, a researcher can preserve their data permanently with the URL of Google Sheet. Data management is also very advantageous because it collects structured data in a single format. KISTI GSDC has been sharing the status of internal clusters for users using Google Sheets. Also, we can distinguish users by Google accounts on the web. We can make the data public to anyone. However, this method requires the consideration of information security. In particular, there need to be in-depth discussions on whether to make it public. It is also necessary to discuss whether it is appropriate to share data which were produced by Korean public institutions to Google which is a foreign company.

Keywords: Google Sheets, homepage-free, monitoring, computing cluster,

1 Introduction

There are many different programs for monitoring computing clusters. Most monitoring programs, such as Ganglia [1] and Elastic Search [2], provide the results in the form of web pages. Most admins used the monitoring system as an internal webpage. However, External users are not authorized to access these internal homepages. Therefore, additional homepages had to be operated by information providers for external users.

In Korea, it is very difficult for information providers to create a homepage with external access to provide very little information to outside users due to the guideline of administration and public institution website [3]. The guideline was legislated to prevent leaking of personal information due to vulnerable sites. The rules have been

[*] Corresponding author

defined to prevent the establishment of meaningless sites and to reinforce the functionality of existing homepages. According to the guidelines, building a new homepage requires a lot of review for security and its needs must be clear.

A website allows users to access and collect data as much as they need. So, information providers can automatically process user's requests. One way to automatically resolve users' requests without running a home page is to provide service APIs. By providing a public API [4], users can acquire the information they want at the time they want.

However, separate programs should be developed to use open APIs. In addition, a program library may be needed to draw charts or graphs to help users understand data easily. This can be very burdensome for the service provider.

We recommend using the Google API to resolve this issue. Users can easily convert data from Google sheets into charts or graphs with simple manipulation. Therefore, the information provider does not have to respond to create a new chart.

2 Related Work

Google Docs is a word processing service on the web [5]. Most of research about Google docs, it was often used as a shared note [6] for researching or a learning tool for students [7], and there is no published content as replacement of homepage for information providing service. This is presumably because there is no statute limiting the website in other countries. However, we were able to find a blog about how to use Google Sheet like a database [8].

3 Concept, Model and Methodology

3.1 Concept of the Google Sheets method

In general, to provide open APIs by informants, the following procedures are performed [9][10].

1. The data-source server stores the generated data in files or database.
2. The contents of saved files or internal database can be accessible from an API server.
3. The API server provides data to external users in the form of a public API or its own API.
4. Users use the data using specific apps or client programs.

Like the Open API method, the Google Sheet method requires informers to collect information from data sources and process it in the form of databases or text files. This processed data is accessible from the Upload Server. Unlike the API servers in Open APIs, Google Sheets' upload server do not need to receive external user's requests directly. Therefore, the upload server itself does not have to allow external access; it simply has to be Internet-enabled to upload data to Google Sheets. The upload server uploads processed data to the Google sheet through the Google API. Unlike API servers with bi-directional communication, the upload server simply forwards information in one direction. Data uploaded to the Google Sheet can be handled by the user. Users can process their data with features such as tables, graphs, and charts on Google sheets.

This method allows service providers to delegate user account information management. If the data manager wants to open data only to specific users, the open API approach should be directly responsible for user authentication and user information management. However, the Google Sheet method minimizes this inconvenience by having the data manager commission user authentication in conjunction with Google account information. Users can also use the service conveniently because there is no additional membership process to use the service.

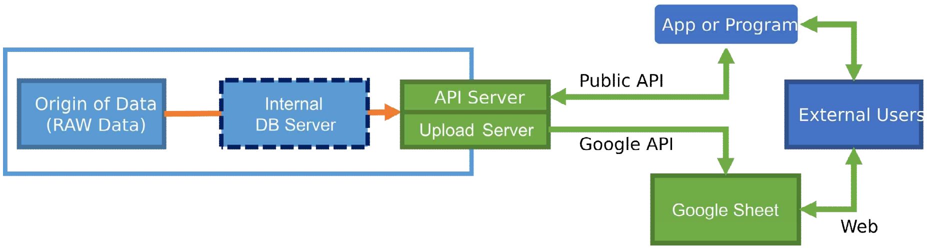


Fig. 1. The API server that provides the Public API requires the data provider to respond to a user's request, but Google Docs using the Google API responds to the user's request from Google and the data provider only needs to upload the data to google.

4 Case Study

4.1 Status of KISTI-GSDC Korea CMS Storage

KISTI-GSDC has operated a storage service for scientists involved in CMS experiment [11]. Because that storage should be shared with about 100 users, we need to understand about the data usage of individual users. To manage data usage, we continuously monitor the usage space. Every time storage is full, we've asked users who are using the most storage to delete their data. User's data usage was obtained manually before requesting. As a result, it was difficult for users to get information about their usage. Since the information did not be provided at the file system level, we had to use the du command in Linux to acquire it. It took several hours to get usage information. Users had to know more about data usage. If that information is provided, the community can have its management of the use of the data. Also, since it is self-regulated without the intervention of the service manager, measures are taken more quickly or prevented in advance.

The storage is divided into three areas:

1. *Dataset:* Dataset space is the space where externally generated experimental or simulation data is stored. Users can request data transfer to use the data. They can also ask us to delete data that they no longer need. Modification of data is not allowed. These transferred data can be grouped according to certain conditions and these bound data are called datasets.

2. *User:* User space is the place where data is processed or extracted by the user based on the data stored in the dataset space. Data is matched to individual users and is not usually shared by multiple users. That is, it can be viewed as private space.

3. *Group:* Group space is a space where many users can share and use data. Group data extracted using data sets is stored in this space and shared with users. Users have created several documents to organize this shared data.

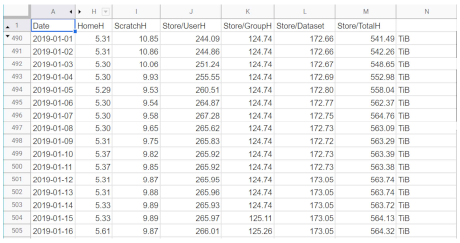
We didn't want to simply divide these three spaces into quotas because we wanted to be flexible. So, changes in data usage in three areas should be monitored for storage management. The dataset space is an area exclusively managed by the service manager, but the user space and group space are managed by individual users or research groups. By analyzing the percentage of use of this storage space, it can be used as meaningful information in establishing storage policies.

4.2 Setup procedure of Google Sheet method

First, we have a server that has access to all three spaces. Administrators have access to each space through directory access on this server. Therefore, we were able to navigate through each space using the Linux du and df commands. We saved the collected information in a text file format. We created a file according to each space. We've written each space file to indicate how much data is being used, based on the highest directory. Also, we have developed a program to communicate the contents of this text file to Google Sheets. We needed the Google Sheets Python API to write the program. Therefore, we installed the Python library on that server to use the Google Sheets Python API by referring to the installation manual provided by Google [12]. We have registered with the crontab service to start the program at dawn every day.

When we first developed the program, we wrote that the program read the spreadsheet document and found the last line. However, we found that the program did not properly handle spreadsheets with more than 1,000 lines. This is because the API was limited to instantaneous throughput. Google Sheet has provided API function that automatically finds the last line to solve this problem. We solved the problem by using *append* functions [13].

Data uploaded to Google Seats can be used as an Excel program from Microsoft through an Internet browser. Drawing a simple chart would be very easy for beginners, too.



	A	H	I	J	K	L	M	N
1	Date	HomeH	ScratchH	Store/UserH	Store/GroupH	Store/Dataset	Store/TotalH	
490	2019-01-01	5.31	10.85	244.09	124.74	172.66	541.49	TiB
491	2019-01-02	5.31	10.86	244.86	124.74	172.66	542.26	TiB
492	2019-01-03	5.30	10.06	251.24	124.74	172.67	548.65	TiB
493	2019-01-04	5.30	9.93	255.55	124.74	172.69	552.98	TiB
494	2019-01-05	5.29	9.53	260.51	124.74	172.80	558.04	TiB
495	2019-01-06	5.30	9.54	264.87	124.74	172.77	562.37	TiB
496	2019-01-07	5.30	9.58	267.28	124.74	172.75	564.76	TiB
497	2019-01-08	5.30	9.65	265.62	124.74	172.73	563.09	TiB
498	2019-01-09	5.31	9.75	265.83	124.74	172.72	563.29	TiB
499	2019-01-10	5.37	9.82	265.92	124.74	172.73	563.39	TiB
500	2019-01-11	5.37	9.85	265.92	124.74	172.73	563.38	TiB
501	2019-01-12	5.31	9.87	265.95	124.74	173.05	563.74	TiB
502	2019-01-13	5.31	9.88	265.96	124.74	173.05	563.74	TiB
503	2019-01-14	5.33	9.89	265.93	124.74	173.05	563.72	TiB
504	2019-01-15	5.33	9.89	265.97	125.11	173.05	564.13	TiB
505	2019-01-16	5.61	9.87	266.01	125.26	173.05	564.32	TiB

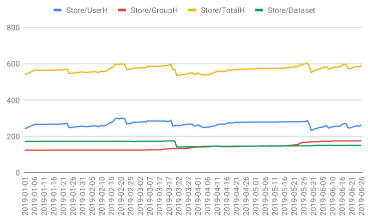


Fig. 2. A simple chart can be drawn using uploaded data in the form of a spreadsheet on a Google sheet. By using this, users can produce a chart or an image as needed.

5 Limitation

The reason why the Google Sheets method is more convenient than running a homepage is that you don't need to manage user account information and server directly. All security issues other than personal information and server management should be managed by the information provider just as they were on the website. In other words, we need to review the contents of information, such as whether it is safe to disclose it.

We assume that this method requires minimal security to comply with public API standards because government guidance did not exist. Likewise, there are expected to be various opinions on whether it is right to provide information produced by public institutions using Google, a foreign company. Of course, the best way is to create an organization that acts on behalf of these services.

6 Conclusion

Through the Google Sheets API, we were able to monitor the data usage of a research group in the GSDC every day. We could easily create charts using data stored in spreadsheet format at Google Sheets.

7 Future Work

The first priority is to make sure that the method is really safe. It's a way of looking at academic possibilities. It requires many discussions to provide real information. If the review is successful, you can apply the Google Sheet method to a wider variety of areas.

8 Acknowledgement

This study was supported by the National Research Foundation of Korea (NRF) through contract N-19-NM-CR01 and the Program of Construction and Operation for Large-scale Science Data Center (K-19-L02-C03).

References

1. Ganglia: Ganglia Monitoring System, http://ganglia.info/.
2. Elastic: Elasticsearch, https://www.elastic.co/kr/products/elasticsearch.
3. Ministry of the Interior and Safety, Administration Public institution web site construction operation guide, https://www.mois.go.kr
4. Open API, https://en.wikipedia.org/wiki/Open_API
5. About Google Docs, https://www.google.com/intl/ko_KR/docs/about/
6. Dekeyser, S., & Watson, R.: Extending google docs to collaborate on research papers. University of Southern Queensland, Australia, 23, (2008)
7. Zhou, W., Simpson, E., & Domizi, D. P.: Google Docs in an out-of-class collaborative writing activity. International Journal of Teaching and Learning in Higher Education, 24(3), 359-375 (2012).
8. The complete guide on how to use Google sheets as a database, https://codingislove.com/google-sheets-database/
9. OpenAPI Initiative, The OpenAPI Specification, https://www. openapis.org/ [Online; accessed February 28, 2018]
10. Hill, G. A., Arrington, T. L., & Hosp, A. K.: *U.S. Patent No. 8,707,276*. Washington, DC: U.S. Patent and Trademark Office., (2014)
11. What is CMS?, http://cms.web.cern.ch/news/what-cms
12. Google Sheets API, https://developers.google.com/sheets/api/
13. Using the API, https://developers.google.com/docs/api/

HTCondor Integrated Cluster Efficiency Analysis of Multiple Research Group through User Job History

Byungyun Kong[1], Sangwook Bae[1], Heejun Yoon[1*] and Seo-Young Noh[2]

[1] Korea Institute of Science and Technology Information (KISTI), 245, Daehak-ro,
Yuseong-gu, Daejeon, Republic of Korea
{kong91, wookie, k2}@kisti.re.kr
[2] Chungbuk National University, Chungdae-ro 1, Seowon-Gu, Cheongju, Chungbuk 28644,
Republic of Korea
rsyoung@cbnu.ac.kr

Abstract. Today, big data analysis requires clusters, not workstations, because it requires a large amount of computing resources. Job scheduling is being introduced to efficiently utilize clusters that are shared by multiple research groups. Job Scheduler has a variety of software such as PBSPro, LSF, SLURM, etc. But we uses HTCondor, which is open source and still active in development and community. While operating these clusters, there is a limitation of resource expansion, but the demand of the users continues, and therefore, a method of increasing the efficiency of existing clusters has been considered. Previously, several groups had their own clusters, which we will try to integrate them physically and divide them into software. To do this, we assigned some work allocation policies and looked at how well the integrated cluster have improved on the basis of user job history.

Keywords: HTCondor, Integrated Cluster, Efficiency, Multiple Research Group

1 Introduction

In a cluster, a batch program helps to efficiently utilize resources through job scheduling. This scheduling is done in various ways such as FIFO and ROUND ROBIN depending on the type of work, and these algorithms suggest ways to get the whole work done faster in the same resource. The clusters we operate are the same. It uses a batch system called HTCondor, which is basically similar to the fair share algorithm[1]. This scheduling helps to efficiently utilize resources in limited resources.

All of the above is a way to efficiently utilize the same resources. However, we currently support multiple research groups, each using its own individual cluster. The clusters used by the users are different from each other according to the research groups.[2] Although a certain resource is allocated, it does not always require all of that resource. Thus, each cluster is operated internally efficiently, but not from a system-wide point of view. One way to solve this problem is to allocate resources flexibly through a container environment.[3] However, because of the difficulty of constructing a container environment for this purpose, we want to integrate the

[*] Corresponding author

clusters through HTCondor's configuration and use a method in which the deployment system allocates resources flexibly.

2 Related Works

HTCondor[4] was developed by the University of Wisconsin research team and has been in service since 1988, known as Condor, but has been using the name HTCondor since 2012. The HTCondor development motivation was developed in order to effectively utilize the power of computing that has been idle and discovers that workstations in each laboratory in the department are more idle than the time they spend most of their time. However, because each researcher or research group that owns the workstation has ownership of computing resources, the range or requirements of the computing resources allowed by each owner are different. In order to effectively solve the situation where the total amount of available computing resources is variable in such a situation, ClassAd is introduced as a kind of job openings/ job hunts to find out cases where the requirements of providers and users are matched with each other, Based scheduling policy.

3 Concept & Plan

Of the several user groups we support, two groups use the same experimental facility[5]. For this reason, two analytic clusters using similar software were selected for integration. The two analytic clusters are similar in their analysis work characteristics, and the work environment configuration and the slot setting such as the resource requirement per job were relatively easy. As an integrated cluster was constructed, the management machine was shortened and the resources for analysis were expanded. As a management policy[6], we guarantee the allocation of computing resources to existing research groups and increase the resource utilization efficiency by applying flexible resource allocation policy. If it is a separate feature, OS dependency is solved by using Singularity due to difference of OS version before integration.

After about 1.5 months of functional testing such as group quota allocation, static slot allocation, dynamic slot allocation, and DAGMan, and after a user beta test such as load test and monitoring code operation check for 1.5 months.

4 Application

The service was officially opened in last November and the official service was provided for about 7 months. The following policies are applied during this period. Resources that are not being used can be used in any group without limit, increasing the efficiency of idle resources. In order to guarantee the use of the existing allocated resources among the integrated resources, if there is an operation request while using another group quota, the most recent operation is canceled and reassigned. This is a policy that is applied considering that the job with the shortest execution time is queued when there is no additional quota. Finally, the allocation priority is applied according to the slot memory size to increase the efficiency of the dynamic assignment machine. Through these policies, we think that resource utilization will be higher than before.

5 Results

We analyze the resource usage from January to May this year during the operation of the integrated cluster and compare it with the resource utilization of the same period last year. Compared to the number of jobs, Group A grew by 137.96%, while Group B grew only by 5.91%. Compared with the waiting time per wall time, group A was shortened by 53.98% and group B was shortened by 15.36%.

Table 1. Resource utilization status of Group A from January to May 2018 and from January to May 2019.

	18’ 1 – 18’ 5	19’ 1 – 19’ 5
Job Count	152,779	363,558
Total Wall Time(s)	384,025,305	1,960,571,192
Total Waiting Time(s)	780,620,180	1,834,174,209

Table 2. Resource utilization status of Group B from January to May 2018 and from January to May 2019.

	18’ 1 – 18’ 5	19’ 1 – 19’ 5
Job Count	2,088,881	2,212,288
Total Wall Time(s)	3,524,414,330	4,907,092,090
Total Waiting Time(s)	6,668,228,020	7,857,942,848

The difference in the efficiency change between the two groups is basically due to the difference in the scale of integration. Prior to consolidation, group A was allocated 400 cores, group B was allocated 1656 cores, and consolidation clusters could aggregate up to 2136 cores. Therefore, when the partner group is not using the resource, the group A acquires 434% of the resources, but only the group B can acquire the 29%.

6 Conclusion

Although the efficiency gains of the two groups differ markedly, this is due to differences in integration scale, both of which have improved absolute efficiency. Therefore, it is expected that the efficiency of the integrated farm will increase as the total size of the various groups is increased. In addition, by predicting the ending time of the work, if the waiting time is short, the work will not be preempted, and if the additional flocking function is used, the utilization of the palm will be higher. However, since there is a risk that all services will be shut down due to a single management server, the service should be stabilized through virtualization or duplication.

Acknowledgments. This study was supported by the National Research Foundation of Korea (NRF) through contract N-19-NM-CR01-S01 and the Program of

Construction and Operation for Large-scale Science Data Center (K-19-L02-C03-S01).

References

1. Basney, Jim, and Miron Livny.: Managing network resources in Condor. High-Performance Distributed Computing, 2000. Proceedings. The Ninth International Symposium on. IEEE, (2000)
2. GSDC Resource Review Board (2018)
3. Bockelman, B., Caballero Bejar, J., & Hover, J.: Interfacing HTCondor-CE with OpenStack, Journal of Physics: Conference Series, vol. 898, pp. 092021 (2017)
4. Miron Livny, Jim Basney, Rajesh Raman, and Todd Tannenbaum,: Mechanisms for High Throughput Computing, SPEEDUP Journal, Vol. 11, No. 1, June (1997)
5. B.Y. Kong, H.J. Yoon, H.J. Han,: Deployment of the Physical Analysis Farm with Workflow Management Application in Data Center, Platform Technology Letters, vol. 4, no. 2, pp. 21-23, (2017)
6. S.U. Ahn, A Jaikar, B.Y. Kong, I Yeo, S Bae and J Kim,: Experience on HTCondor batch system for HEP and other research fields at KISTI-GSDC, Journal of Physics: Conference Series, vol. 898, pp.082013, (2017)

Study of LIGO Tier-2 System with Priorities

Sangwook Bae[1], Geonmo Ryu[2], Seo-Young Noh[3] and Heejun Yoon[4]*

[1,2,4] Korea Institute of Science and Technology Information (KISTI), 245, Daehak-ro,
Yuseong-gu, Daejeon, Republic of Korea
{wookie[1], geonmo[2], k2[4]}@kisti.re.kr
[3] Chungbuk National University, Chungdae-ro 1, Seowon-Gu, Cheongju, Chungbuk 28644,
Republic of Korea
rsyoung@cbnu.ac.kr[3]

Abstract. The emergence of data-driven research and development is perceived as a new paradigm shift. This is based on the production, processing, and analysis of large volumes of data produced by large research facilities and research equipment. As a result, the role of the data center in storing, processing, and analyzing large amounts of data is important in recent years. Global Science Experimental Data Hub Center (GSDC) of Korea Institute of Science and Technology Information (KISTI) is a data center established by a national funding project to promote fundamental research activities in South Korea. GSDC is currently serving as Asia's leading data center in line with the changes in the new research paradigm. Especially, the first direct observation of gravitational waves for the first time in 2015 led to an increase in demand for gravitational wave research, which has led to a rise in demand for global expansion(Tier2 service) of the GSDC-LIGO Data Grid(LDG) system. Therefore, this paper proposes and tests ways to reflect existing users and global requirements of the LIGO Tier3 computing resources of the currently operating GSDC.

Keywords: Data Center, Global Service, Tier2, LIGO Data Grid, HTCondor CE

1 Introduction

Global Science Experimental Data Hub Center (GSDC) of the Korea Institute of Science and Technology Information (KISTI) has established a gravitational wave data analysis computing environment (LDG) to support domestic and foreign researchers[1-3]. GSDC participated in the initial planning stage where KGWG discussed joining LIGO and began to provide full-fledged service to the LDG from 2010. Today, the computing environment of GSDC-LDG is building and operating 996 Core HTCondor-based deployment systems[4], 550TB of dedicated storage and data analysis pipeline software. With the increasing demand for computing for gravitational wave research, more and more people are calling for global conversion of resources currently used exclusively for KGWG and KAGRA users. Therefore, if current resources are

[*] Corresponding author

expanded globally, the existing user and grid job must be accommodated. In this paper, we propose and test a scheme to efficiently operate existing user jobs and grid jobs through HTCondor CE and HTCondor (Local CE). The remainder of this paper is organized as the follows: Chapter 2 describes the status of the GSDC-LDG, identifies the considerations for Layer 2 services, and applies and tests. The paper concludes with Chapter 3.

2 Study of LIGO Tier-2 System with Priorities

For existing GSDC-LDG, 996 cores, 550TB are supported and configured with HTCondor deployment systems to provide services. Recently, CVMFS(CernVM File System) has also been serviced, supporting LIGO-related data and SW to users. It also has about 220 TB of LIGO data and about 15 TB of KAGRA data. In addition, the GSDC-LDG is registered with the LIGO central monitoring system, and LIGO data is sent from the California Institute of Technology in the United States to the GSDC via the data transfer system, the LIGO Data Replicator (LDR) server. To extend these systems to Global Services, LIGO Tier2, HTCondor CE must be applied to existing systems. HTCondor CE is a special configuration of HTCondor software designed as a working gateway solution for LIGO grid Job. For LIGO grid Job, HTCondor CE is composed of 8.6.13 version, and HTCondor for local scheduling is composed of 8.8.3 version supporting dynamic slot. In addition to existing KGWG and KAGRA user jobs, the following should be considered for the efficient operation of LIGO grid operations. First, for OSG Grid jobs, consideration for existing users is required because resources are used almost as full as supported. Second, by default, the job of KGWG users should have high priority, except for the dedicated resources for LIGO grid jobs. Third, for KAGRA users, there is not always a Job, so efficient operation of idle resources is required.

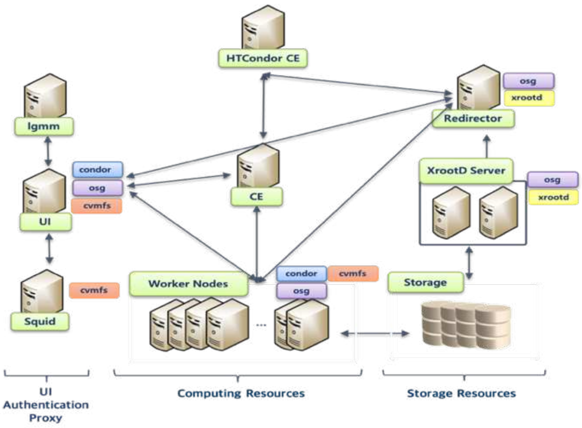


Fig. 1. The GSDC-LDG system for global expansion is Computing Elements, Storage Elements and ETC. Consist of Computing Element was built primarily as an HTCondor batch system and was built on the basis of Storage Elements Xrootd.

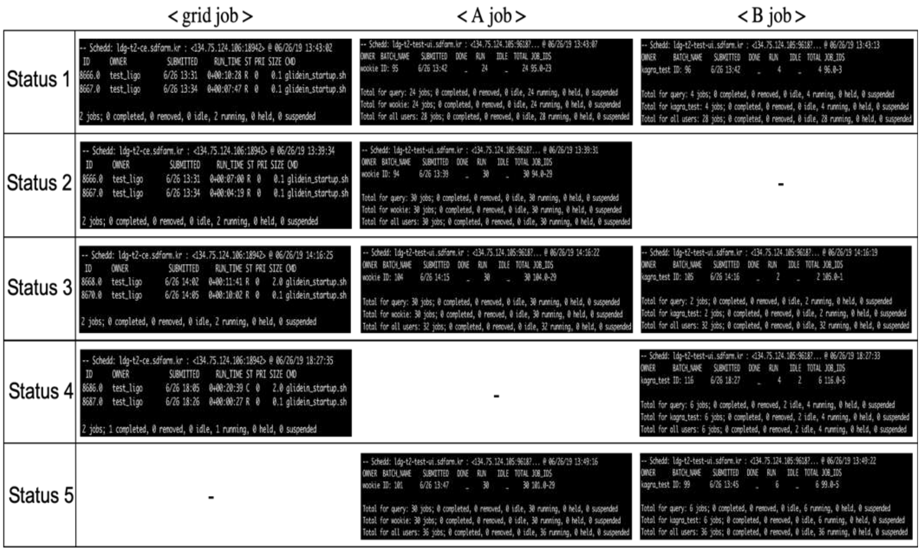
Table 1. The environment for testing consisted of 9 servers in total

Role		Cores	Memory	Quantity
Storage	Xrootd Redirect Server	12 core	24GB	1
Element	Xrootd Server	36 core	168GB	3
	UI	12 core	24GB	1
Computing	HTCondor CE	12 core	24GB	1
Element	HTCondor WN	48 core	144GB	2
	Local CE	12 core	24GB	1

In addition, settings that reflect considerations should be set up in the local CE and UI for GROUP and PREEMPT as shown in Fig. 2. Also, settings that reflect considerations should be added to local CE settings for HTCondor's GROUP and PREEMPT. The environment for testing is 20 cores assigned to group grid, 24 cores assigned to group A, and 4 cores assigned to group B:

```
GROUP_NAMES = group_grid, group_A, group_B
GROUP_QUOTA_group_grid = 20
GROUP_ACCEPT_SURPLUS_group_grid = false
GROUP_QUOTA_group_A = 24
GROUP_ACCEPT_SURPLUS_group_A = true
GROUP_QUOTA_group_B = 4
GROUP_ACCEPT_SURPLUS_group_B = false
```

The GROUP and PREEMPT settings that reflect considerations were set up and tested according to the circumstances as follows: First, if grid job, A job and B job are 16, 24 and 4 cores respectively, it can be seen that each core is running correctly(Fig. 2, Status 1). Second, if grid job, A Job, and B Job are 16, 30, and 0 cores respectively, the A Job is used in addition to the resources allocated to it. Therefore, it can be seen that all 30 cores requested are working(Fig. 2, Status 2). Third, if grid job, A Job, and B Job are 16, 30, and 2 cores respectively, the A Job uses all resources other than those allocated to it. Therefore, it can be seen that all 30 cores and 2 cores requested are working(Fig. 2, Status 3). Also, If grid job, A Job, and B Job are 16, 0, and 6 cores respectively, only 4 of B job can be executed even if they have resources, and the other 2 jobs can be confirmed to be processed after 4 jobs are completed(Fig. 2, Status 4). Finally, if grid job, A Job, and B Job are 0, 30, and 6 cores respectively, both request A Job and B Job are running because grid job is empty(Fig. 2, Status 5).



< grid job > < A job > < B job >

Status 1

```
-- Schedd: ldg-t2-ce.sdfarm.kr : <134.75.124.106:18942> @ 06/26/19 13:43:02
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
8666.0   test_ligo       6/26 13:31   0+00:10:28 R  0    0.1 glidein_startup.sh
8667.0   test_ligo       6/26 13:34   0+00:07:47 R  0    0.1 glidein_startup.sh

2 jobs; 0 completed, 0 removed, 0 idle, 2 running, 0 held, 0 suspended
```

```
-- Schedd: ldg-t2-test-ui.sdfarm.kr : <134.75.124.105:9618?... @ 06/26/19 13:43:07
OWNER  BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
wookie ID: 95       6/26 13:42     _     24     _      24 95.0-23

Total for query: 24 jobs; 0 completed, 0 removed, 0 idle, 24 running, 0 held, 0 suspended
Total for wookie: 24 jobs; 0 completed, 0 removed, 0 idle, 24 running, 0 held, 0 suspended
Total for all users: 28 jobs; 0 completed, 0 removed, 0 idle, 28 running, 0 held, 0 suspended
```

```
-- Schedd: ldg-t2-test-ui.sdfarm.kr : <134.75.124.105:9618?... @ 06/26/19 13:43:13
OWNER      BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
kagra_test ID: 96       6/26 13:42     _      4     _       4 96.0-3

Total for query: 4 jobs; 0 completed, 0 removed, 0 idle, 4 running, 0 held, 0 suspended
Total for kagra_test: 4 jobs; 0 completed, 0 removed, 0 idle, 4 running, 0 held, 0 suspended
Total for all users: 28 jobs; 0 completed, 0 removed, 0 idle, 28 running, 0 held, 0 suspended
```

Status 2

```
-- Schedd: ldg-t2-ce.sdfarm.kr : <134.75.124.106:18942> @ 06/26/19 13:39:34
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
8666.0   test_ligo       6/26 13:31   0+00:07:00 R  0    0.1 glidein_startup.sh
8667.0   test_ligo       6/26 13:34   0+00:04:19 R  0    0.1 glidein_startup.sh

2 jobs; 0 completed, 0 removed, 0 idle, 2 running, 0 held, 0 suspended
```

```
-- Schedd: ldg-t2-test-ui.sdfarm.kr : <134.75.124.105:9618?... @ 06/26/19 13:39:31
OWNER  BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
wookie ID: 94       6/26 13:39     _     30     _      30 94.0-29

Total for query: 30 jobs; 0 completed, 0 removed, 0 idle, 30 running, 0 held, 0 suspended
Total for wookie: 30 jobs; 0 completed, 0 removed, 0 idle, 30 running, 0 held, 0 suspended
Total for all users: 30 jobs; 0 completed, 0 removed, 0 idle, 30 running, 0 held, 0 suspended
```

-

Status 3

```
-- Schedd: ldg-t2-ce.sdfarm.kr : <134.75.124.106:18942> @ 06/26/19 14:16:25
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
8668.0   test_ligo       6/26 14:02   0+00:11:41 R  0    2.0 glidein_startup.sh
8670.0   test_ligo       6/26 14:05   0+00:10:02 R  0    0.1 glidein_startup.sh

2 jobs; 0 completed, 0 removed, 0 idle, 2 running, 0 held, 0 suspended
```

```
-- Schedd: ldg-t2-test-ui.sdfarm.kr : <134.75.124.105:9618?... @ 06/26/19 14:16:22
OWNER  BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
wookie ID: 104      6/26 14:15     _     30     _      30 104.0-29

Total for query: 30 jobs; 0 completed, 0 removed, 0 idle, 30 running, 0 held, 0 suspended
Total for wookie: 30 jobs; 0 completed, 0 removed, 0 idle, 30 running, 0 held, 0 suspended
Total for all users: 32 jobs; 0 completed, 0 removed, 0 idle, 32 running, 0 held, 0 suspended
```

```
-- Schedd: ldg-t2-test-ui.sdfarm.kr : <134.75.124.105:9618?... @ 06/26/19 14:16:19
OWNER      BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
kagra_test ID: 105      6/26 14:16     _      2     _       2 105.0-1

Total for query: 2 jobs; 0 completed, 0 removed, 0 idle, 2 running, 0 held, 0 suspended
Total for kagra_test: 2 jobs; 0 completed, 0 removed, 0 idle, 2 running, 0 held, 0 suspended
Total for all users: 32 jobs; 0 completed, 0 removed, 0 idle, 32 running, 0 held, 0 suspended
```

Status 4

```
-- Schedd: ldg-t2-ce.sdfarm.kr : <134.75.124.106:18942> @ 06/26/19 18:27:35
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
8686.0   test_ligo       6/26 18:05   0+00:20:39 C  0    2.0 glidein_startup.sh
8687.0   test_ligo       6/26 18:26   0+00:00:27 R  0    0.1 glidein_startup.sh

2 jobs; 1 completed, 0 removed, 0 idle, 1 running, 0 held, 0 suspended
```

-

```
-- Schedd: ldg-t2-test-ui.sdfarm.kr : <134.75.124.105:9618?... @ 06/26/19 18:27:33
OWNER      BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
kagra_test ID: 116      6/26 18:27     _      4      2      6 116.0-5

Total for query: 6 jobs; 0 completed, 0 removed, 2 idle, 4 running, 0 held, 0 suspended
Total for kagra_test: 6 jobs; 0 completed, 0 removed, 2 idle, 4 running, 0 held, 0 suspended
Total for all users: 6 jobs; 0 completed, 0 removed, 2 idle, 4 running, 0 held, 0 suspended
```

Status 5

-

```
-- Schedd: ldg-t2-test-ui.sdfarm.kr : <134.75.124.105:9618?... @ 06/26/19 13:49:16
OWNER  BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
wookie ID: 101      6/26 13:47     _     30     _      30 101.0-29

Total for query: 30 jobs; 0 completed, 0 removed, 0 idle, 30 running, 0 held, 0 suspended
Total for wookie: 30 jobs; 0 completed, 0 removed, 0 idle, 30 running, 0 held, 0 suspended
Total for all users: 36 jobs; 0 completed, 0 removed, 0 idle, 36 running, 0 held, 0 suspended
```

```
-- Schedd: ldg-t2-test-ui.sdfarm.kr : <134.75.124.105:9618?... @ 06/26/19 13:49:22
OWNER      BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
kagra_test ID: 99       6/26 13:45     _      6     _       6 99.0-5

Total for query: 6 jobs; 0 completed, 0 removed, 0 idle, 6 running, 0 held, 0 suspended
Total for kagra_test: 6 jobs; 0 completed, 0 removed, 0 idle, 6 running, 0 held, 0 suspended
Total for all users: 36 jobs; 0 completed, 0 removed, 0 idle, 36 running, 0 held, 0 suspended
```

Fig. 2. GROUP and PREEMPT have been set up and tested to reflect considerations in various situations.

3 Conclusion and Future Work

The role of the data center has become important due to recent changes in the research paradigm. The GSDC is working to serve as Asia's leading data center. Especially, the first direct observation of gravitational waves for the first time in 2015 led to an increase in demand for gravitational wave research, which has led to a rise in demand for global expansion of the GSDC-LDG system. Therefore, there is a need to maintain the convenience of current KGWG users and KAGRA users, while providing stable service to global jobs. In this paper, the system environment was established for this purpose and tested to reflect considerations. These results enable expansion into stable global services in the future.

References

1. S U Ahn, A Jaikar, B Kong, I Yeo, S Bae and J Kim.: Experience on HTCondor batch system for HEP and other research fields at KISTI-GSDC. Journal of Physics: Conference Series, Conf. Series 182938, 2017.
2. Korea Institute of Science and Technology Information https://www.en.kisti.re.kr/
3. LIGO https://www.ligo.caltech.edu/.
4. HTCondor Version v8.8 Manual http://research.cs.wisc.edu/htcondor/manual/v8.8.
5. Geonmo Ryu, Seoyoung Noh.: Establishment of new WLCG Tier Center using HTCondorCE on UMD middleware. CHEP 2018 Conference, Sofia, Bulgaria, 2018.

Survey on Network Data Transfer Systems and Research Data Move Requirements

Jin Kim[1] , Seo-Young Noh[2] , Hee Jun Yoon[1*]

[1] Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu,
Daejeon, Korea
{jkim, k2}@kisti.re.kr
[2] Department of Computer Science, Chungbuk National University, Chungdae-ro 1,
Seowon-Gu, 28644, Cheongju, Korea
rsyoung@cbnu.ac.kr

Abstract. The atmosphere that dramatic evolution of computing and networking is going together is certainly clear because the size of data which is used for scientific discovery or experimental research is getting larger. In this paper, we focus on the networking element to share the experiment data between computing facilities and network transfer systems for large scientific data.

Keywords: network tuning, scientific bulk data transfer, DTN

1 Introduction

It is a clear that the size of research data is getting bigger and bigger. As a same manner, the size of computing, networking, storage system is also too. To get the large size of data, the huge scale research equipment what is used to find a novel scientific discovery is constructed in various research fields. After that equipment generate the scientific data, then that should be used in large computing facility, shared many researchers, and stored massive archiving storage. At each share step, it is the most important consideration that network features such as network bandwidth, delay, jitter on network device (router, switch) and window size, buffer size, parameter optimization, data transfer middleware on server. Nowadays, the development direction of those network features is different as what is the main purpose of data transfer. In the case of normal data such as Internet traffic, the QoS of data transfer is depend on transfer middleware such as FTP, Torrent, etc., because of the difficult to set up the network device parameter and to predict the data routing path. As a result of that, it is not a necessary consideration that normal user who use normal data worry about network QoS, because the normal network backbone bandwidth is large enough to use normal data and the QoS is enough [1]. In the case of scientific-experimental data, here are several characters of the data is that first is the size of experimental raw data is large, second is generating of the data is not easy

[*] Corresponding Author

because it needs large scale of experimental equipment, third is data generates other meaningful data through each research level (that is why scientific-experimental data is large then normal data), forth is it should be shared for common purpose what is spectacular achievement in each research area. At the first season of computing, there are separated parts (computer HW, SW, database, software engineering, AI, network, and etc.) and developed in each direction. However, the data science which data is central in experimental research must be considered every computing element, specially connectivity between each computing element. In this paper, we talk about the atmosphere of large scale computing and networking evolution and in various computing part and which kind of consideration is mentioned to make a better system with network view because all parts of computing element should be connected.

2 Background

2.1 Science DMZ and Data Transfer Node (DTN)

In the 2000s, the increase of grid computing use and the birth of cloud computing were an opportunity to see the importance of networking technology and also to see the limitation of network development without server consideration. Science DMZ model is the most innovative network model since 2010. The main idea of Science DMZ model is to divide user traffic and scientific traffic on network [2]. The advantage of doing this is: first, user traffic cannot disturb the scientific traffic transfer which consumes a lot of time on the network, second, the scientific traffic does not pass the security device, third, the network bandwidth can be guaranteed for large scientific data transfer. The idea of science DMZ is from the character of scientific data which is the size of it is too large and the network requirement such as bandwidth, delay, security is important to transfer that data. Data Transfer Node is the dedicated system to transfer the data. There are 4 DTN workflows for DTN designing; storage type, networking protocol, motherboard on subsystem, operating system [3].

2.2 SENSE (SDN for End-to-End Networked Science at the Exascale) project

Recently, the one of most active project what is SENSE (SDN for End-to-End Networked Science at the Exascale) project [4]. It is building smart network services for data intensive computing. The basic technology is SDN but there are a lot of network elements are considered in SENSE project. First is unified resource orchestration between distributed infrastructure. SDN technology can divided the traffic depend on use or character, however, it is not easy to control SDN technology on different network environment. SENSE Orchestrator gathers network situation to make a SENSE End-to-End Model using SENSE-RM API in different network domain and DTNs.

3 Data Transfer System – BigData Express Project

Recently, Data Transfer System is collectively called all related element includes server, network, OS, NIC, storage, middleware, etc. because it is not enough to consider single element for better performance. In this section, we would like to introduce transfer middleware. BigData Express project [5] is to address common and fundamental problems for bulk data transfer in the extreme-scale era. First problem is that existing data transfer middle ware is not enough to cover various elements (network status, server condition and so on.) to transfer the data. Second problem is that existing solution do not support to minimize cross interference between data transfer. Third problem is that it is hard to satisfy the user requirement such as deadline transfer. Last problem is that existing middleware is not fit to run on DTN node because the middleware does not consider the specification of server or operating system. The main goal of the project is to build a distributed middleware system that will provide a schedulable, predictable, and high performance data transfer service for the large-scale science facilities and their collaborators. There are 4 major components on BDE; web portal as a user interface, scheduler for the resource management and resource orchestration, mdtmFTP is a high performance data transfer engine, amoebaNET for SDN technology.

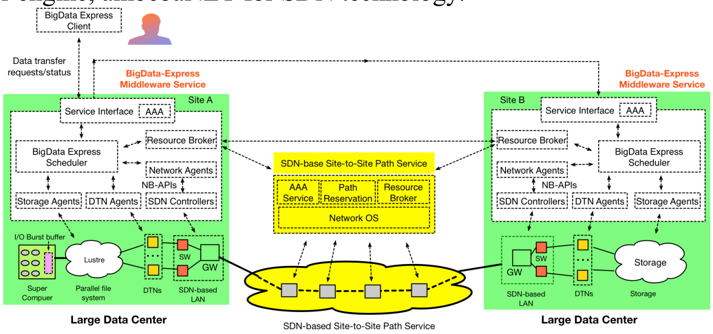


Figure 1. BigData Express architecture

DTN Agents manage data transfer nodes (DTNs). This is the component of Big Data Express that is responsible for: 1. Querying network and storage configuration and capabilities of the DTN. 2. Registering itself with the database, thus advertising the availability, configuration, and capabilities of the DTN to BDE Server. 3. Launching, monitoring, and re-starting the data transfer application (such as MdtmFTP or GridFTP) as necessary. 4. Initializing Storage Agents, based on configuration. 5. Acting upon, and responding to, commands from BDE Server. 6. Setting up the data transfer node with job-specific configurations, such as adding virtual circuits or routes, and verifying that the configuration works. 7. Monitoring and rate control data transfer. 8. Tearing down job-specific configuration once the job is complete.

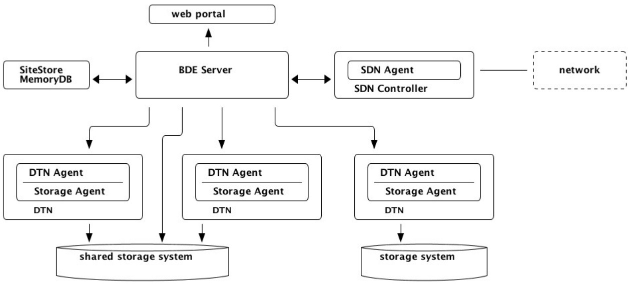


Figure 2. DTN 에이전트 구조

4 Conclusion

Nowadays, scientific data transfer between complex computing systems are essential requirement. We suggest a new point a view to construct data transfer system design. To transfer the scientific bulk data, the necessary system requirements are the fast storage system, high performance DTN(HW) and middleware(SW), software awareness network service.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) through contract N-19-NM-CR01-S01 and the Program of Construction and Operation for Large-scale Science Data Center (K-19-L02-C03-S01).

References

1. network service QoS evaluation report by Ministry of Science and ICT, https://www.msit.go.kr/web/msipContents/contentsView.do?cateId=mssw311&artId=1461858, 2018
2. Eli Dart, Lauren Rotman, Brian Tierney, Mary Hester, Jason Zurawski: The Science DMZ: A Network Design Pattern for Data-Intensive Science: hindawi.com, vol 22, issue 2, pp. 173-185 (2014)
3. Eric Pouyoul: design and build your data transfer node: Internet2 tutorial, (2012)
4. Inder Monga, Chin Guok, John Macauley, Alex Sim, Harvey Newman, Justa Balcas, Phil DeMar, Linda Winkler, Tom Lehman, Xi Yang: SND for End-to-End Networked Science at the Exascale (SENSE): IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS), pp. 33-44, (2018)
5. Qiming Lu, Liang Zhang, S.Sasidharan, Wenji Wu, Phil DeMar, Se-young Yu, Jim Hao Chen, Joe Mambretti, Xi Yang, Tom Lehman, Chin Guok, J. Macauley, Inder Monga, Jin Kim, Seo-Young Noh: BigData Express: Toward Schedulable, Predictable, and High-Performance Data Transfer; Innovating the Network for Data Intensive Science workshop, Supercomputing Conference 2018, (2018)

Analysis of Daily Traffic on Global Science Experimental Data Hub Center in Korea

Jin Kim[1], Hee Jun Yoon[1], Jinsoo Park[2,*]

[1]Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu,
Daejeon, Korea
{jkim, k2}@kisti.re.kr
[2]Department of Management Informations Systems, Yong In University, 134 Yingindaehak-ro, Cheoin-gu, Yongin, Korea
jsf001@yongin.ac.kr

Abstract. This study analyzes daily traffic on Global Science experimental Data hub Center in Korea. Since the collected traffic volume data is imperfect, we refine the data through preprocessing. The 7-day moving average of daily traffic volume makes the original data into proper time series data. With the refined data, we examine the self-similarity and estimate the Hurst parameter through rescaled range test. From the analysis and its results, we will draw some limitations and significant implications.

Keywords: GSDC, daily traffic, time series analysis, self-similarity

1 Introduction

Global Science experimental Data hub Center (GSDC) is a data center that supports basic sciences operated by Korea Institute of Science and Technology Information (KISTI). The network traffic is a remarkable factor to provide stable service. Analysis of GSDC network traffic will facilitate establishment of strategies for load balancing or introduction of new equipment. Therefore, this study analyzes the GSDC network traffic through time series approach. The network topology of GSDC is spine-leaf structure with 2 backbone spine switches and a lot of leaf switches are located on the top-of-rack [1]. The bandwidth between spine and leaf switches is 80Gbps and the virtual router redundancy protocol is enable on the spine switch 80Gbps. The connection to the storage also has same capacity. KISTI has gathered the network in/out bound traffic for two types of switch. One is the internal traffic that goes to storage side, and the other is external traffic that goes to user connection or other servers. In this research, the collected traffic is time series traffic data of the daily uplink traffic volume. In generally, it is known that the time series data of World Wide Web (WWW) traffic or complex network traffic has self-similarity [2], [3]. Accordingly, this study tries to investigate the self-similarity of the target traffic from GSDC datacenter. In the data collection, however, there are lots of mistakes and

[*] Corresponding Author

omissions due to the collection procedure of the traffic volumes because operator manually collect the data. Since this situation makes the data imperfect, there is a difficulty in the time series analysis for daily network traffic volume. Another difficulty in the analysis is arisen from the spine-leaf structure of GSDC. Those two phenomena will also be discussed.

This paper is organized as follows. Section 2 describes the traffic data collection procedure and explains the data preprocessing to refine the imperfect data into a proper time series data. Section 3 analyzes the time series data to examine the self-similarity, and draws some discussions. Section 4 presents conclusions and future works.

2 Traffic Data Collection from GSDC network

2.1 Mistakes and Omissions in Data Collection Procedure

In GSDC, data center network traffic data is collected manually. When a collector requests the daily volume data to the system, the system responds the accumulated traffic volumes of respective switches. Then he or she copies the data to the clipboard, and collects to a spreadsheet file by pasting them. There are two possible incidents in this collection procedure. In the step of copy and paste, some mistakes can arise from the collector or the result string. After an equipment is serviced, its traffic volume data becomes initialized. The response can produce irregular results. In addition, the absence of collector on weekends and holidays causes the omission of the data. These phenomena make the daily data to an imperfect time series data. Therefore, this study refines the collected traffic data into proper time series data.

2.2 Data Preprocessing

To refine the imperfect daily data into a proper time series data, we preprocess the collected data. Since the data is accumulated values, we calculate the daily traffic by differencing the successive data. In this step, we manually review all the data and consult the person in charge to identify the initialized data and the mistakes. Although the daily traffic data is calculated, it cannot be treated as a time series data. There are still missing data on weekends and holidays. To make a proper time series data, we calculate the 7-day moving average (MA) for daily traffic using the accumulated data. The MV data is not perfect but is analyzable significant time series data. Therefore, this paper does not mark the date values at the date axis on all the graphs.

This study generates the MA data using the collected data from Jan. 10 to Jun. 28 in 2019. GSDC supports seven experimental groups; ALICE, CMS, LIGO, HCP, TEM, BIO, and RENO. On the GSDC network structure, respective group is connected to one or more network switches. Accordingly, we aggregate the traffic volumes for respective group. This paper shows the results from ALICE group since it is one of the representative group that generates the largest traffic volume.

3 Time Series Approach

3.1 Moving Averaged Data

For respective groups, there are four types of traffic volumes; internal inbound (in-in), internal outbound (in-out), external inbound (ex-in), and external outbound (ex-out). Fig. 1 plots the responded original traffic data and the refined MA data for those four types from switches connected to ALICE group server. As shown in Fig. 1 (a), there are missing and abnormal data in the original data. Fig. 1 (b), however, shows the proper shapes as time series data. Since in-out traffic volume has low scale compare with others, its plot in Fig. 1 (b) seems like straight line. We notify that the data has its own pattern.

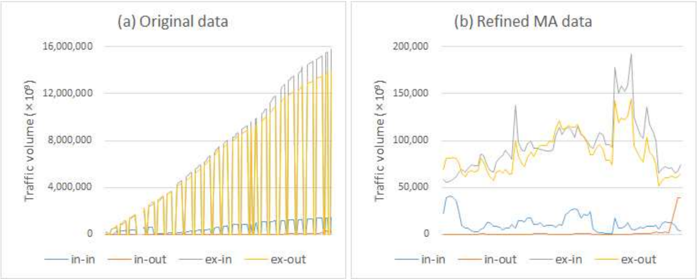


Fig. 1. Responded original data and refined MA data.

3.2 Examining Self-similarity

Self-similarity is well known concept in network traffic. In general, there are four methods to assess self-similarity; the rescaled range (or R/S) plot, the variance-time plot, the period gram plot, and the Whittle estimator. For simplicity and ease to understand, we apply R/S plot to examine the self-similarity. For a detailed procedure and notations of R/S plot, see [2] and [4].

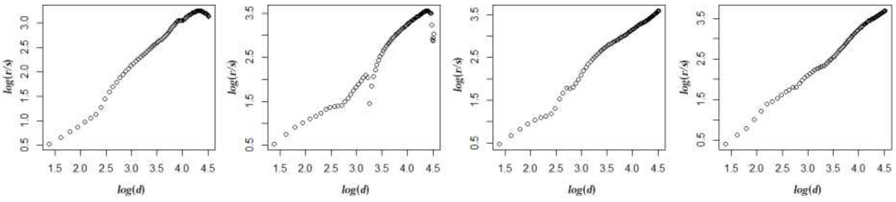


Fig. 2. Rescaled range (R/S) plots for four types of traffic volumes.

Fig. 2 shows the possibility that the MA data has self-similarity, and the values of Hurst parameter, *H* are estimated as 0.7336, 0.7541, 0.7686, and 0.7794 for respective types.

3.3 Discussion

There are some limitations in this study. Since we collected daily traffic for about six months, data points are insufficient to explain the self-similarity. As the second limitation, we only test the R/S plot. To complete the self-similarity examination, other three tests are also required.

From this study, we also draw some significant implications. The data collection environment is remarkably inadequate considering GSDC scale and KISTI retained technology. It is also important to evaluate the network performance as the traffics have self-similarity. It will be helpful the load balancing, and establishment of strategies to introduce new equipment.

4 Conclusion

We analyzed the daily network traffic of GSDC with time series approach. We refined the imperfect original data into proper time series data. Analysis shows that the daily traffic estimated to have self-similarity even though the result is incomplete. We also drew that the data collection system must be established for proper performance evaluation.

Since this study has some limitations, we need to supplement them in the future. In addition, this study only shows the results for ALICE group server. Analysis of all the experimental group and their comparison will be a fruitful research topic.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) through contract N-19-NM-CR01-S01 and the Program of Construction and Operation for Large-scale Science Data Center (K-19-L02-C03-S01).

References

1. Kim, J., Yeo, I., Cho, K.: Data Center Network Characteristic Analysis through a GSDC Case Study. Platform Technology Letters, 4(2), 2014--2017 (2017)
2. Crovella, M.E., Bestavros, A.: Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes, IEEE/ACM Transactions on networking, 5(6), 835-846 (1997)
3. Song, C., Havlin, S., Makse, H.A.: Self-similarity of complex networks, Nature, 433(7024), 392 (2005)
4. Willinger, W., Taqqu, M.S., Leland, W.E., Wilson, D.V.: Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements. Statistical science, 10(1), 67—85 (1995)

Topic analysis of Internet news articles

Kyoung-ae Jang [1,1], Misun Park[1,2], Woo-Je Kim [1,3],

[1] *Dept. of Industry and Information Systems Engineering, Seoul National Univ. of Science and Technology,*
232, Gongneung-ro, Nowon-gu, Seoul, 01811, Republic of Korea
{ Frontier Laboratory Room 701}

Abstract. We gathered recent internet news articles and analyzed the meaning of what is being discussed regarding the news articles. News articles were basic data suitable for analyzing the fourth industrial revolution. We used the LDA and word2vec analysis method. We collected internet news articles from 2017 to 2018 and analyzed the main topics with LDA. Word2Vec analysis was performed to extract the sub-keyword. This study reveals concepts and directions of the fourth industrial revolution.

Keywords: the fourth industrial revolution, Latent Dirichlet Allocation analysis, Word2Vec analysis, Topic modeling

1 Introduction

We live in the early stages of the fourth industrial revolution era. We have a lot of expectations and interests in the future of the fourth industry. We began this research to know the direction of the fourth industrial age that previous researchers expected. The first industrial revolution began with the advent of the 18th century steam engine. The second industrial revolution appeared with the computers, and the third industrial revolution appeared through the Internet. Now, the fourth industrial revolution has started with the advent of IOT(Internet of Things) based artificial intelligence [1][2].

Currently, industry and academia are aware of the importance of the fourth industrial revolution and investing actively in the industrial revolution. However, the study of the fourth industrial revolution is still in low level [2]. Therefore, we gathered the latest internet news articles and analyzed the meaning of what is being discussed regarding the fourth industrial revolution. News articles are suitable for analyzing the contents of the fourth industrial revolution.

This study reveals concepts and directions of the fourth industrial revolution in business, government, and industry by topic analysis for the fourth industrial revolution. This study can help guide relevant future policy and research on the fourth industrial revolution.

[1] First Author
[2] Second Author
[3] Corresponding Author

2 Previous Research

2.1 LDA(Latent Dirichlet Allocation)

We analyzed the distribution of word counts to predict document subjects using latent dirichlet allocation (LDA). The LDA is a hierarchical bayesian model proposed by Blei et. al. to classify words and build topic models [3]. The LDA is an unsupervised learning method that probabilistically analyzes a document topic using the dirichlet distribution [4]. It basically uses latent semantic indexing (LSI) to find hidden topics in documents, and it solved overfitting problem of LSI [5][6].

2.2 Word2Vec

Mikolov et. al. proposed word2vec as an unsupervised learning method which computes the meaning of a word as a vector of values based on a neural network technique [7][8]. Word2vec is described by the skip-gram model and can predict words within the specified range around the input word.

The closer the distribution of words is, the more likely it is represented as a similar vector. The studies for word2vec are diverse such as the study of word clustering by Pyysalo et. al. and the study of recognition about biomedical group by Tang et. al., and so on [9][10].

3 Research Design & Results

This study is comprised of the following steps: First we collected data including internet news data from 2017 to 2018. Next step is preprocessing to remove stop words, special characters, and space. Then we accomplish LDA analysis using matrix analysis models incorporating words, article names, nouns, and keywords. Next step is word2vec analysis for each topic by year of publication.

We collected internet news data from 2017 to 2018 using web crawling with the keyword “the fourth industrial revolution”. Then Korean morphological analysis was performed using open source KoNLPy[11] to drive morpheme from the collected internet news in preprocessing step. In addition, we removed stop-words and derived nouns in preprocessing methods which were proposed by Hylth and You who showed that keywords tended to be included in nouns [13].

We analyzed the articles by year of publication of the internet news. LDA analysis was performed with the refined articles by year of publication. The results of LDA analysis were classified into four topics with a high of with-class scatter. The results of the LDA analysis in 2017 articles included the presidential election, the president, the people and the government in Topic 1, and corporate, investment, artificial intelligence, AI and autonomous driving in Topic 2. In Topic 3, education, college, students, and subject were organized. In Topic 4, government, policy, and reinforcement were formed. Word2Vec analysis was performed to extract the sub-keywords for top 4 keywords which were elicited by the LDA analysis. The subordinate keywords were derived by cosine similarity from the keywords defined in the LDA topic groups.

4 Conclusion

The purpose of this study was to classify subjects related to the fourth industrial revolution concept and interest. We collected internet news articles from 2017 to 2018 and analyzed the main topics. We first performed LDA analysis to derive main topics and keywords. As a result of LDA analysis, the topics were politics, policy, education, and new technology in 2017; in 2018 main topics related to the fourth industrial revolution were policy, new technology, industry and finance. Word2Vec analysis was performed to extract the sub-keywords by the LDA analysis.

This study can provide an important basis for understanding key issues of the fourth industrial revolution in academia and industry in Korea. The public interests and core issues change constantly as shown by the topic analyses over the research periods. Thus, continuous research is required to track these changes as the fourth industrial revolution progresses.

References

[1] J. Bloem, M. Van Doorn, S. Duivestein, D. Excoffier, "The Fourth Industrial Revolution", sogeti.com., 2014.

[2] K. Schwab, "The fourth industrial revolution", Crown Business, 2017.

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation”, Journal of machine Learning research, pp.993-1022, 2003.

[4] Frigyik, Bela A., Amol Kapila, and Maya R. Gupta. "Introduction to the Dirichlet distribution and related processes.", Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006, 2010.

[5] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. "Indexing by latent semantic analysis", Journal of the American society for information science, vol.41, no.6, pp.391-407, 1990.

[6] Thomas Hormann, "Probabilistic latent semantic indexing", In Proceedings of the 22nd annual international ACM SIFIR conference on Research and development in information reterieval, pp.50-57, 1999.

[7] Mikolov, T., Chen, K., Carrado, G., and Dean, J., "Efficient estimation of word representations in vector space". arXiv preprint arXiv:1301.3781, 2013,

[8] Mikolov, T., Lee, Q. V., and Sutskever, I., "Exploiting similarities among languages for machine translate", arXiv preprint arXiv:1309,4168, 2013,

[9] Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S., "Distributional Semantics Resources for Biomedical Text Processing", LBM 2013 proceeding, 2013.

[10] Tang, B., Cao, H., Wang, X., Chen, Q., and Xu, H., "Evaluating word representation features inf biomedical named entity recognition tasks". BioMod research international, 2014.

[11] http://konlpy.org/en/v0.4.4/data, 2018.

[13] W. You, D. Fontaine, and J. P. Barthes, "An automatic keyphrase extraction system for scienrific documents.", Knowledge and information systems, vol.34, no.3, pp.691-724, 2013.

[14] K. Schwab, “The fourth industrial revolution”, Crown Business. 2017.

[15] S. Klaus, “The fourth industrial revolution”, In World Economic Forum, pp.11, 2016.

[16] Kyoung-ae Jang, Misun Park1, Woo-Je Kim, “Topic Analysis for the fourth Industrial Revolution using LDA in Korea”, International Journal of Engineering, Construction and Computing (IJECC), 2018.

Pre-training of Deep Bidirectional Transformer for Text Clustering

Tien Anh Nguyen[1] , Hyung-Jeong Yang *[1]
[1] Department of Electronics and Computer Engineering, Chonnam National University,
Gwangju, South Korea
*Corresponding author: *hjyang@jnu.ac.kr*

Abstract. Text clustering is attracted a lot of attention of researchers and widely applied to numerous applications in unsupervised machine learning. There are a lot of approach to improve text clustering performance from two main perspectives: feature extraction and distance function. In this study, we try to adapt and extend performance of deep embedding clustering model for text clustering by applying Deep Bidirectional Transformer and convolutional operation. We challenge our propose model to short and long text clustering. The experiments demonstrate that our propose model show significant improvement over reference methods.

Keywords: unsupervised learning, deep learning, text clustering.

1 Introduction

Text clustering is central research challenge in unsupervised learning research field. Researchers spend effort into discovery and propose new approaches to improve clustering performance from different perspectives: similarity measurement or distance metric, feature extraction and so on. To develop an effective clustering model, we need to do two steps: (1) convert input data to features space, (2) measure the similarity of feature data points. However, can we provide a clustering model which has ability to learn optimal parameters for feature extraction and dissimilarity measurement by applying breakthrough of deep learning. The first breakthrough solution is come from Junyuan [1]. The authors propose a deep learning model for both text and images clustering. They use Kullback-Leibler (KL) as a loss function to provide the model ability to learn optimal parameters for feature extraction and similarity measurement automatically. In this study, we try to adapt and extend the performance of model for specific text clustering problem. In this study, we try to capture the meaning of text data input effectively by using pre-train Deep Bidirectional Transformer (BERT) [2] model to capture contextual meaning of text. Then, we apply convolutional operation to extract lower level feature for clustering. We challenge our model with two clustering problems: short and long text clustering. We also provide a comparison of our method to other ones.

2 Related works

In 2015, Junyuan [1] introduce a new approach which is called deep embedding clustering (DEC) that tries to use data driven to resolve clustering challenge. The model uses autoencoder to convert input data to features space. Then, it applies the Student's t-distribution as a kernel to measure the similarity between data points. It defines a new loss function based on Kullback-Leibler (KL) divergence. The method quite high performance for both images and text data. However, autoencoder cannot provide embedding features for words based on their context. In other words, after the training process, autoencoder provides an output as a dictionary which contains embedding for all words. Each word has the same embedding in every sentence and context.

In 2017, Jiaming [3] propose a new approach which is called Short Text Clustering via Convolutional neural networks (STCC) for short text clustering. The author uses Locality Preserving Indexing binary codes as supervison for convolutional operation extract lower level features from word embedding. Then, those features are transfer to K-mean to do clustering. However, K-mean is a traditional machine learning algorithm. We cannot apply data driven approach to K-mean to resolve the clustering challenge.

3 Proposed method

In this section, we describe our proposed method which is developed from Deep Bidirectional Transformer (BERT) [2], convolutional layers [4] and deep embedding clustering. In [1], the authors use deep autoencoder to convert text input to feature vector space. Then, the Student's t-distribution is applied to compute the dissimilarity between data point for clustering. The formula of Student's t-distribution is represented in the formula (1):

$$q_{ij} = \frac{\left(1 + \left\| z_i - \mu_j \right\|^2\right)^{-1}}{\sum_{j'}(1 + \left\| z_i - \mu_{j'} \right\|^2)^{-1}} \qquad (1)$$

However, the most important disadvantage is that output of autoencoder model is a dictionary which contains unique representation of every word. It means that each word has the same representation in all sentence and context. To resolve this drawback, we proposed to use BERT to extract contextual embedding features of text document input. Unlikely autoencoder, BERT is a deep learning model takes text document as input and produce embedding features for each word in this document. Therefore, each word will have difference representation in difference context. BERT has two version: BERT BASE and BERT LARGE. BERT BASE uses a vector which has the size of 768 to describe embedding features, whereas BERT LARGE compacts embedding into a vector of 1024. In this study, we recommend BERT LARGE for extracting contextual word embedding features. Therefore, a sentence of length N has a numerical form:

$$X_{1:N} = x_1 \oplus x_2 \oplus x_3 \oplus \dots \oplus x_n \qquad (2)$$

Where ⊕ is concatenation operator. In other words, a sentence has a N words is represented as a matrix N×1024.

Then, we apply convolution operation to this matrix to extract lower level features of words. Base on experimental results, we adapt convolution architecture which is developed by Kim [5] for this purpose. The operation involves 3 kernel sizes: $5 \times 1024, 4 \times 1024, 3 \times 1024$. We scan the matrix features $X_{1:N}$ by 100 filters for each kernel sizes. After that, all the outputs of those kernel sizes are concatenated together and flatten to convert to 2 fully connected layers (FC) which have the same size of 1024. The output is now ready for clustering.

To clustering, we use clustering layer which is implemented by the formula (1) [1]. Intuitively, this formula tries to assign probability data point $z_i$to centroid μ_j, where z_i is embedding data point of document input $X_{1:N}$. Similarity to [1], we still apply Kullback-Leibler (KL) divergence to define loss function. This loss function is defined as the formula (3):

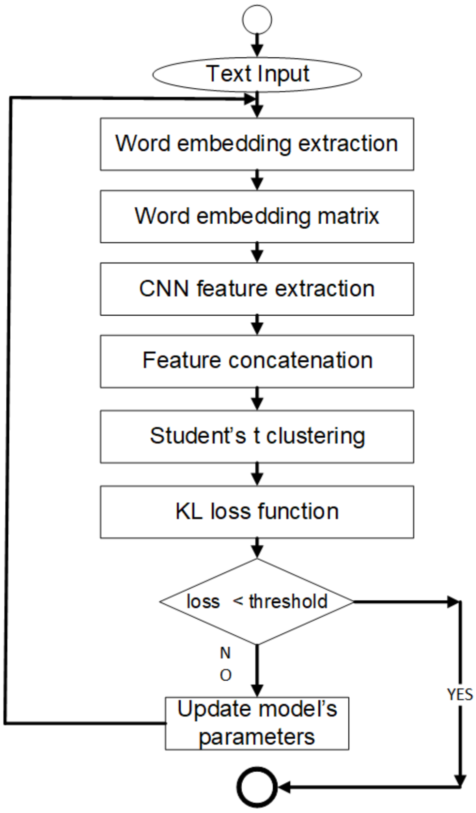


Figure 1. Proposed Model's Architecture

$$L = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{3}$$

where $p_{ij} = \frac{q_{ij}^2/f_i}{\sum_{j'} q_{ij'}^2/f_{j'}}$ and $f_j = \sum_i q_{ij}$. The general architecture is described in the Figure 1.

4 Experimental results

4.1 Datasets

In this study, we challenge our proposed model to long text and short text challenge. In the first trial, we apply this model to Reuters [7] dataset. The Reuters dataset contains more than 800000 English stories. They are very long documents. Because pre-trained BERT model only adapt document of maximum 512 words, we fix maximum length for document is 512 and remove the rest word of the documents. Similar to DEC [1], we use four root categories: corporate/industrial (CCAT), government/social (GCAT), markets (MCAT) and economics (ECAT) as single ground truth of the dataset.

In the second challenge, we test the model's ability with Stackoverflow [2] dataset. The Stackoverflow dataset is contained 20000 sentences which are focus on some specific topics of Information Technologies (IT) collected from Stackoverflow website. The labels of this dataset is described in the Table 1. We also summarize general information of three datasets in the Table 2.

Table 1. Labels of Stackoverflow dataset

svn	oracle	bash	apache	excel
matlab	cocoa	visual-studio	osx	Wordpress
spring	hibernate	scala	sharepoint	ajax
drupal	qt	haskell	linq	magento

Table 2. Datasets Description

Datasets	Data size	Type
Reuters (RCV1)	800,000 documents	Long text
Stackoverflow	20,000	Short Text

4.2 Evaluation metric

Similarity to DEC, we use unsupervised clustering accuracy (ACC) which is described in the formula (4) as an evaluation method.

$$ACC = \max_m \frac{\sum_{i=1}^{n} 1\{l_i = m(c_i)\}}{n} \tag{4}$$

where l_i is the ground-truth labels and c_i is the cluster outputs of proposed model and m is all possible one-to-one mappings between l_i and c_i. Therefore, we can apply Hungarian algorithm to find the best mapping between c_i and l_i to calculate ACC.

4.3 Evaluation and results

In this section, we compare our proposed model to DEC and STCC in long and short text challenge, respectively. The results are described in the Table 3 and 4. In the Table 3, we compare our proposed model with K-mean and DEC models. According to experimental results, our proposed model outperformance 26.55% and 4.21% than K-mean and DEC models, respectively. In the Table 4, we compare our model with STCC in short text challenge by applying both models to Stackoverflow dataset. The proposed model overs 1.16% than STCC method.

Table 3. Long Text Clustering Experiments

Methods	**Reuters**
K-mean	53.29%
DEC [1]	75.63%
Proposed model	79.84%

Table 4. Short Text Clustering Experiments

Methods	**Stackoverflow**
STCC [2]	51.14%
Proposed model	52.3%

5 Conclusion

In this study, we try to adapt and improve DEC model for text clustering by applied pre-train BERT model. We choose BERT large version for our study. BERT is applied to extract contextual word embedding of text input. Then, we applied convolutional operation to extract lower level features of embeddings. We still using clustering layers of DEC for text clustering. Our model outperformance than reference methods. In the future, we will try to modify convolutional architecture to extract features more effectively to achieve better clustering performance.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2019-2016-0-00314) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation).

References

1. Xie, J., Girshick, R., & Farhadi, A.: Unsupervised Deep Embedding for Clustering Analysis. In M. F. B. and K. Q. Weinberger (Ed.), Proceedings of Machine Learning Research, New York, New York, USA (2016)
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv Preprint ArXiv:1810.04805 (2018)
3. Xu, J., Xu, B., Wang, P., Zheng, S., Tian, G., Zhao, J., & Xu, B.: Self-Taught convolutional neural networks for short text clustering. Neural Networks, vol. 88, pp. 22--31. Elsevier (2017)
4. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE, vol. 86, pp. 2278--2323, IEEE (1998)
5. Kim, Y.: Convolutional Neural Networks for Sentence Classification. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1746--1751, Doha, Qatar (2014)
6. Carr, C. E., Khutsishvili, I., & Marky, L. A.: CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. AAAI Conference on Web and Social Media, vol. 122, pp. 7057--7065, Palo Alto, California, US (2018)
7. Lewis, D. D., Yang, Y. M., Rose, T. G., & Li, F.: RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, vol. 5, pp. 361--397, (2004)

Fake News Detection Framework Using Word Networks and Propagation Patterns of News in Social Media

Jongmo Kim, Hanmin Kim, Jinuk Cho and Mye Sohn[1]

Department of Industrial Engineering, Sungkyunkwan University,
Suwon, Korea
{dignityc, kimhm0705, brbl, myesohn}@skku.edu

Abstract. The fallacy of fake news is becoming more and more critical due to the characteristics of social media such as low cost, rapid propagation and echo chamber effects. Although there are several attempts to detect fake news using the machine learning techniques, they cannot detect fake news that is time-critical and sophisticatedly written. To overcome the limitations, we proposed the novel framework to detect the fake news using the word networks and propagation patterns of fake news in social media. We devise the calculation of the weights based on $\mathrm{tf}-\mathrm{idf}$ to modify the word networks and the propagation patterns generation method applied the LDA-based topic modeling.

Keywords: Fake News Detection, Word Networks, LDA, News Propagation Pattern

1 Introduction

Due to the characteristics of social media such as low cost, easy accessibility, rapid propagation, and the echo chamber effect, fake news with intent to deceive users is flooding into social media [1]. These characteristics of social media not only promote mass production of the fake news but also worsen the negative influence of them [2]. Many researchers have been attempted fake news detection using machine learning techniques based on the news contents [3]. However, existing research has critical limitations. First, fake news is typically time-critical, but content-based methods cannot detect it in a timely manner. Second, because sophisticatedly written fake news is very similar to true news, it is very difficult to detect fake news with only the information it contains. To overcome the limitations, this paper used “social engagements” for detecting fake news on social media. At this time, social commitments are defined as user behaviors that express opinions in various forms, such as sharing, commenting, and posting [7].

To detect fake news considering social engagements, we propose a novel framework that performs two types of classification learning for different purposes: the word co-occurrence networks-based contents classification and propagation patterns-based classification with social engagements. The former is conducted to detect the truly true news through analysis of news contents with textual features. To do so, we use the word networks of the co-occurrence among words that can find the implicit relationships

[1] Corresponding author

among the words and calculate the weight of the words. The word networks are used as the input data for abstaining learning that is appropriate for the imbalance dataset. Finally, the learner conducts classification and removal the truly true news from the fake news dataset. The latter is performed to detect fake news from the news dataset that is removed the truly true news. To do so, it uses propagation patterns of the news attaching social engagements. The propagation patterns of the news for the specific topic are identified by the Latent Dirichlet Allocation (LDA) topic modeling. As a next, the ensemble learning using the social engagements as the features is conducted to obtain a robust classifier for various propagation patterns of fake news.

This paper is organized as follow. In Section 2, we review the research of fake news detection. Section 3 offers the overall fake news detection framework using word networks and propagation patterns and their detailed process. Section 4 demonstrate the superiority of the proposed framework. Finally, Section 5 presents the conclusions and further research.

2 Related Works

Research to detect fake news can be divided into two categories that use news contents and social engagements as the features. In the research using the news contents, there are two approaches: linguistic-based approach [4-6] and data driven approach [8, 9]. The research has limitations that they cannot reflect the news trend and overlook the features of social media, which are important information to analyze patterns of them. To overcome the limitation, some studies have attempted to exploit the social features to detect the fake news in social media such as frequency of likes, number of sharing news, and information of followings as social features [10, 11]. Also, others have utilized the networks theory to interpret the features about social media [12, 13]. However, due to the characteristics of social engagement like noisy and incompleteness, they are struggling to apply the machine learning algorithms with high performance. Clearly, the new approach is needed to solve the distinct characteristics of social engagement data disturbing the machine learning algorithms. The detailed descriptions are summarized in Table 1.

Table 1. Analysis of the fake news detection method

		Descriptions	**Refs.**
News contents	Linguistic -based	- Based on term frequency or Linguistic Inquiry Word Count, linguistic feature such as style and syntax. - Limitation: do not to deal with the fast-changed writing style	[4-6]
	Data driven	- Attempt to discover implicit or explicit features of the news contents using the data mining techniques like LSTM or neural network	[8, 9]
Social engagements	User-based	- Based on each user reaction to the news - Classification of logistics regression	[10]
	Post-based	- Used to features obtained from post sharing news using Bayesian inference	[11]
	Network-based	- Find the features of propagation patterns using CNN and RNN - Using networks formed between users or posts	[12, 13]

3 The Overall of Framework

As depicted in Fig. 1, our framework is divided into two modules that are performed sequentially. First module is word networks-based truly true news classifier learning module. It is performed to reduce the size of the training dataset by removing the majority of the truly true news data from the fake news training dataset._To do so, our framework performs abstaining learning to solve an issue of the data imbalance, and filter-based feature selection for fast learning. Second module executes the propagation-based ensemble learning using the reduced dataset. The purpose of the second module is to learn robust fake news classifier using propagation patterns of the news with social engagements. To do so, we perform LDA-based propagation patterns generation for each topic. In order to convert the propagation pattern as a matrix for learning, we construct the feature set using social engagements. Finally, we conduct the wrapper-based feature selection and ensemble learning to obtain robust classifier for detecting the fake news.

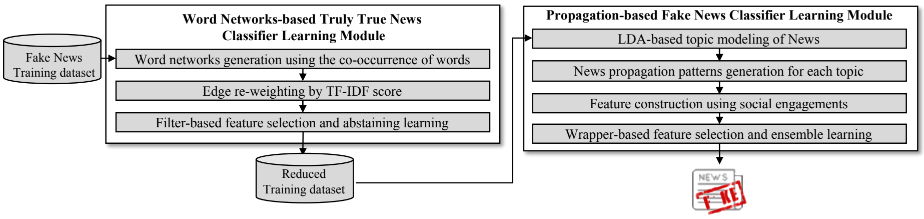

Fake News
Training dataset
Word Networks-based Truly True News
Classifier Learning Module
Word networks generation using the co-occurrence of words
Edge re-weighting by TF-IDF score
Filter-based feature selection and abstaining learning
Reduced
Training dataset
Propagation-based Fake News Classifier Learning Module
LDA-based topic modeling of News
News propagation patterns generation for each topic
Feature construction using social engagements
Wrapper-based feature selection and ensemble learning
NEWS
FAKE

Fig. 1. Fake news Detection framework

3.1 Word Networks-based Truly True News Classifier Learning Module

This module has two objectives. The one is to reduce the volume of the training dataset of the fake news training dataset. The other is to solve the class imbalance problem of the news dataset. Although it is the training dataset of fake news, it is very likely that it contains a lot of truly true news. The truly true news increases the volume of the dataset, which may hinder the speed up the learning rate. In addition, it causes the class imbalance problem to disturb the classifier when it classifies the minority class like fake news. To resolve problems, first, we perform the classifier learning to remove the truly true news from the training dataset of fake news. As a way of classifier learning, we apply the word networks method using the co-occurrence of words, which can find the implicit relationships among the words and give the weights of the words, too.

The word networks are generated using the co-occurrence frequency of two words within a sentence of the news. In traditional network representation, the undirected graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ is represented as a set of V and E where V and E denote the set of vertices and edges. Based on the representation, the word network of each news contents is defined as follows.

Definition 1 i^{th} word networks (N_i) is an undirected graph with a multiset of words (W_i) as vertex and the co-occurrence relationships of the words in sentences as edges. It is simply represented as follows.

$$N_i = (W_i, E_i) \tag{1}$$

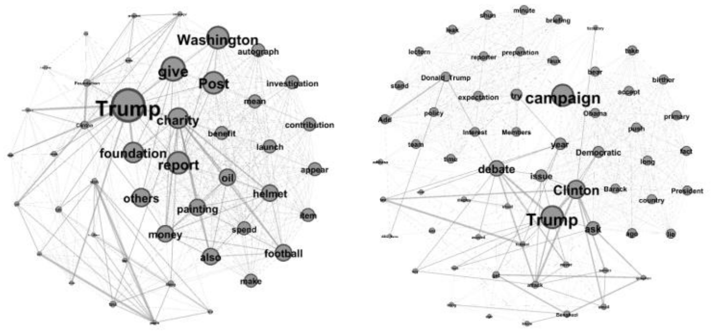
where $W_i \ni \{w_{i1}, w_{i2}, \dots, w_{ij}, \dots\}$and $E_i \ni \{e_{i1}, e_{i2}, \dots, e_{ik}, \dots\}(j, k \in \mathbb{N})$.

Definition 2 k^{th} edge (e_{ik}) of N_i is composed of related two words and its weight (wt_{ik}). It is simply represented as follows.

$$e_{ik} = \{(w_{ij}, w_{ij'}), wt_{ik}\}, ij \neq ij' \tag{2}$$

where wt_{ik} is initially set to the sum of the co-occurrence frequencies of w_{ij} and $w_{ij'}$ in all sentences.

Through the generated word networks, we can grasp not only the distribution of the words but also the relationship between them. To show the capability of the word networks, we created the two types of the word networks using 2016 US presidential election news of the BuzzFace. As a result, we can see that true news and fake news has the word networks with different shapes (Fig. 2). In this light, the word networks are appropriate input data for classifier learning.



The word network of true news The word network of fake news

Fig. 2. Shape of word networks of true news and fake news

As state earlier, wt_{ik} is calculated only by the co-occurrence frequency of the words in the sentences. It causes a problem that it is possible to give a high weights not-informative but frequent common words such as 'others' or 'also' in Fig. 2. To solve the problem, we conducted to update the wt_{ik} using weights of two words, which is derived by $\mathrm{tf-idf}$ method. At this time, $\mathrm{tf-idf}$ of is calculated for term set $T = \{t_1, t_2, \dots, t_l, \dots\}$ that is eliminated duplicate words contained in W $(T \subset W)$. The algorithm to calculate the $\mathrm{tf-idf}$ score of a term t_l is summarized in Fig. 3.

The word networks should be transformed into the feature vectors for performing abstaining learning. To do so, we identified two categories of the features that can best reveal the nature of word networks: network-based features and non-network-based features. The network-based features such as common neighbors, betweenness centrality, the length of the longest path, and core number is used as metrics to estimate the shape of a network. On the contrary, the non-network-based features are basic statistics to summarize the dataset. It contains the number of words and edges, the sum of $\mathrm{tf-idf}$ scores, and the average of edge weights. Using the network-based features and the non-network-based features, we perform the filter-based feature selection. To evaluate the performance of features, we use the concept of mutual information. Also, to solve the imbalance problem, we conducted the abstaining learning using SVM classifier [15]. By the abstaining learning, we obtain the delicate classifier that can reject the classification of fake news. In addition, it can classify the truly true news with the high probability threshold.

```
Input: word networks $N_i$
Output: updated weight $wt_{ik}$

for all $t_l \in T$ do
  generate $tfidf_l$ variable
  Calculate_tfidf($t_l$)
end for
for all $e_k \in N_i$ do
  $wt_{ik} = wt_{ik} \times$ Search_tfidf($w_{ij}$) $\times$ Search_tfidf($w_{ij'}$) and $w_{ij}$, $w_{ij'} \in e_k$
end for
Search_tfidf($w_{ij}$)
  for all $t_l \in T$ do
    if $w_{ij} == t_l$ do; return $tfidf_l$
Calculate_tfidf($t_l$)
  $f_l$ = Count $t_l$ in N
  $fN_l$ = Count $N_i \ni t_l$ in N
  $tfidf_l = (0.5 + \frac{0.5+f_l}{\max(f_l)}) \times log \frac{|N|}{fN_l}$
```

Fig. 3. Update algorithm of Edge weight

3.2 Propagation-based Fake News Classifier Learning Module

In this module, we generate the propagation patterns of the news using LDA topic modeling. Since the shape of propagation patterns are different for each topic and it is not revealed in the dataset, the LDA topic modeling should be performed to discover and generate the propagation patterns of news for specific topics. The process of LDA topic modeling is simply divided two steps: the generative process for a word [16] and Gibbs sampling [17]. After the LDA topic modeling, we can obtain the topic distribution (θ) in news and is defined as follows.

Definition 3 t^{th} topic distribution (θ_t) is a vector of the t^{th} topic probabilities for the all of N_i

$$\theta_t = \{z_{t1}, z_{t2}, \dots, z_{ti}, \dots\}^T \tag{3}$$

where z_{ti} is the t^{th} topic probability for the N_i

Using the topic distribution θ_t, we generate the propagation patterns for the specific topics. First, we find the median for all z_{ti} ($med(z)$). Second, in the topic distribution θ_t, we find the z_{ti} higher than the $med(z)$ and select the N_i corresponding to it.

Next, it generates the propagation patterns using the update time of selected N_i. In this module, we consider only temporal properties of propagation except the structure and linguistics properties, because these properties are partially considered in the word networks. The propagation patterns generation algorithm is illustrated in Fig. 4.

```
Input: word networks $N_i \in N^-$, $N^-$ is reduced dataset by removing the truly true news
Output: propagation patterns

generate $\theta_t$ vectors
perform the LDA topic modeling using $N^-$ return $\theta_t$
generate $P_t$ variables
for all $\theta_t$ do
    for all $z_{ti} \in \theta_t$ do
        if $z_{ti} > med(z)$ then $P_t \xleftarrow{append} updateTime\ of\ N_i$
return $P_t$
```

Fig. 4. Propagation patterns generation algorithm

We construct the feature set based on the propagation P_t and social engagements. The propagation pattern-based features are distinct features only observed from time series data: total population, staring time, background noise, interaction periodicity offset. In addition, we include the social features by social engagements data such as frequency of the likes, number of friends, and number of news sharing.

Finally, we perform the wrapper-based feature selection and ensemble learning to detect the fake news. Although these methods require high computational costs, it is suitable for our framework because the dataset was reduced by first module. Though the learning process, we obtain the robust the classifier to detect the fake news considering propagation patterns and social engagements.

4 Experiments

To prove the performance of framework, we prepared the experimental dataset using the BuzzFace, which is the labelled fake news dataset by comprising the news data from disparate sources. In the BuzzFace dataset, we verified and selected the 895 news data, because some news is vague for determining fake or true and have no news contents data. We preprocessed the news contents data by tokenizing, stemming, lemmatization and removing stop words using the NLTK (Natural Language Toolkit) and conducted resampling the dataset to adjust the ratio of fake news to test the performance for imbalance dataset. Using the preprocessed and resampled dataset, we performed the learning based on SVM and Random Forest (RF) 10 times using the abstaining learning (AL) and non-abstaining learning (NoAL) approach (all classes have same cost). The learning result is shown in Fig. 5.

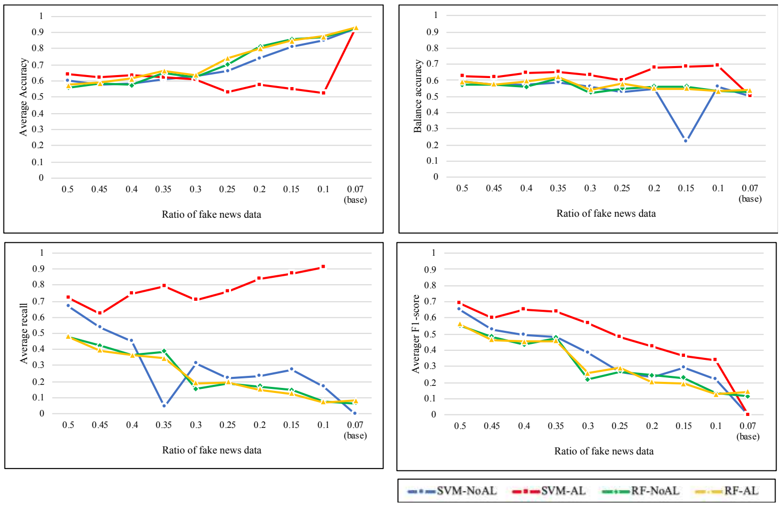


Fig. 5. The learning results using SVM and RF with abstaining learning

To evaluation performance of classification, we utilized four metrics: accuracy, recall, f1-score, and balanced accuracy. Unlike the former three metrics, the balanced accuracy is a suitable metric to evaluation the performance of imbalance dataset classification. In terms of accuracy, other methods seemed to outperform our method (SVM-AL). However, the accuracy metric is inappropriate to evaluate the imbalance dataset. Therefore, in the balance accuracy, which is suitable to evaluate the imbalance dataset, our method outperforms the other methods and have high score according to the severe imbalance dataset in recall. Hence, our method is better to the imbalanced fake news dataset than other method.

5 Conclusion and Future works

We proposed the novel frame work to detect the fake news using the word networks and LDA-based propagation patterns generation methods. It has a significance to mix the different feature sets by the word networks and propagation patterns to handle the characteristics of social media. Also, the framework adopts the different the feature selection methods and classifier learning method to attain both efficiency of accuracy of the fake news detection framework. However, it should be improved for real-time detection and high dimensional data classification.

In the future work, we will adopt the semantic web techniques to improve the contents analysis such as Linked Open Data or semantic latent matrix. In addition, we consider the graph-based algorithm to consider the graph structure of propagation patterns. The deep learning methods (CNN and LSTM) could be considered to improve the performance of machine learning techniques such as opinion mining in social media.

Acknowledgements This research is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2019R1A2C1004102).

References

1. LAZER, David MJ, et al. The science of fake news. Science, 2018, 359.6380: 1094-1096.
2. SHU, Kai, et al. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 2017, 19.1: 22-36.
3. AHMED, Hadeer; TRAORE, Issa; SAAD, Sherif. Detection of online fake news using n-gram analysis and machine learning techniques. In: International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. Springer, Cham, 2017. p. 127-138.
4. RASHKIN, Hannah, et al. Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.
5. RUBIN, Victoria, et al. Fake news or truth? using satirical cues to detect potentially misleading news. Proceedings of the second workshop on computational approaches to deception detection. 2016.
6. PEREZ-ROSAS, Verónica, et al. Automatic detection of fake news. arXiv preprint arXiv:1708.07104. 2017.
7. Khan, M. L. (2017). Social media engagement: What motivates user participation and consumption on YouTube?. Computers in Human Behavior, 66, 236-247.
8. POPAT, Kashyap, et al. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. arXiv preprint arXiv:1809.06416 (2018).
9. SINGHANIA, Sneha, Nigel Fernandez, and Shrisha Rao. 3han: A deep neural network for fake news detection. International Conference on Neural Information Processing. Springer, Cham, 2017.
10. Tacchini, Eugenio, et al. "Some like it hoax: Automated fake news detection in social networks." arXiv preprint arXiv:1704.07506 (2017).
11. Tschiatschek, Sebastian, et al. "Fake news detection in social networks via crowd signals." Companion of the The Web Conference 2018 on The Web Conference 2018. International World Wide Web Conferences Steering Committee, 2018.
12. Conti, Mauro, et al. "It's always April fools' day!: On the difficulty of social network misinformation classification via propagation features." 2017 IEEE Workshop on Information Forensics and Security (WIFS). IEEE, 2017.
13. Liu, Yang, and Yi-Fang Brook Wu. "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
14. CUKIERSKI, William; HAMNER, Benjamin; YANG, Bo. Graph-based features for supervised link prediction. In: The 2011 International Joint Conference on Neural Networks. IEEE, 2011. p. 1237-1244.
15. Zhang, X., & Hu, B. G. (2014). A new strategy of cost-free learning in the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 26(12), 2872-2885.
16. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
17. Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7), 424-440.

18. Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013, December). Prominent features of rumor propagation in online social media. In 2013 IEEE 13th International Conference on Data Mining (pp. 1103-1108). IEEE.
19. Khan, M. L. (2017). Social media engagement: What motivates user participation and consumption on YouTube?. Computers in Human Behavior, 66, 236-247.

Construction of the Database for Providing all Cosmetic Ingredients Safety Information

Jung-hyun Ahn[1], Doorheon Jeong[2], Menghok Heak[2], Sang-Hyun Choi[2],

[1] Department of Big Data, Chungbuk National University,
Cheongju, South Korea
[2] Department of Manangement Information System, Chungbuk National University,
Cheongju, South Korea
{typass21, djccnt15, menghok.heak}@gmail.com, chois@cbnu.ac.kr

Abstract. With the recently happening harmful effects of using baby cosmetics, consumers are increasingly demanding safety information. In order to meet this, various information providing systems such as blogs or applications have appeared, but the problem of data reliability still remains. Therefore, in this study, reliability of all cosmetic ingredients safety information was secured by building a database (DB) of the data of all ingredients of cosmetics safety based on cosmetic raw material regulation information for each country and cosmetic ingredient safety evaluation information provided by Korea Cosmetic Industry Institute (KCII). In order to construct this database, the standardization work has been done by using all cosmetic ingredient name data and list of the standardized names of the ingredients data.

Keywords: Baby Cosmetics; Data Reliability; Raw Material Regulation Information; Ingredient Safety Evaluation Information; Standardization

1 Introduction

Recently, as incidents and accidents related to cosmetics have been frequently occurred, consumers are interested in safety of cosmetic ingredients. In fact, according to one survey, about 78% of survey respondents said they are interested in cosmetic safety [1]. In order to guarantee the consumers' right to know about the safety of cosmetics, the government implemented "The Cosmetics Ingredient Labeling System", which will label all the ingredients used in cosmetics manufacturing in container or packaging starting from October 2008 [2].

However, the cosmetics ingredient labeling system currently makes it difficult for consumers to understand and to get information due to using terminology and simple list of ingredients name [3]. The number of blogs, applications, and SNSs keeps increasing where consumers can not only get all ingredient names of cosmetics but also find and share information related to them [4]. These blogs and applications provide useful information about cosmetic safety from a consumer perspective, but they also provide unproven information. That adds to consumers' confusion about the choice and use of cosmetics [5].

Therefore, in this study, reliability of all cosmetic ingredients safety information was secured by building the database (DB) of the data of all ingredients of cosmetics safety based on cosmetic raw material regulation information for each country and cosmetic ingredient safety evaluation information provided by Korea Cosmetic Industry Institute (KCII). In addition, a visualization web page has been developed based on the database that enables consumers to easily check safety information at a glance. Such research and system development can be a cornerstone of future studies on establishing safety standards for cosmetics, and it is valuable in providing consumer-oriented information that enhances reliability and visibility.

2 Database Construction

To develop a database of all cosmetic ingredients safety information, three types of data mentioned in Table 1 were collected. The first is data of the brand names of cosmetics and the names of all the components. In this study, data were collected focusing on baby cosmetics, one of the hottest issues in recent years. The data for all ingredient names of cosmetics were collected from online shopping malls through crawling. Second, the standardization name for all ingredients of cosmetics was collected by using Product Ingredient Standardized data provided by Korea Cosmetic Association(KCA). The data crawled by the online shopping mall is a mixture of old name and standard name. Finally, we obtained the safety level information of all components of the sample products using KCII DB.

Table 1. Data description

Dataset Name	Dataset ID	Record	Source
Product Ingredient	Set 1	Product type Product name Ingredient name	Online shopping mall
Product Ingredient Standardized	Set 2	Standard name, Standard English name Old name Old English name	Korea Cosmetic Association
Ingredient safety level	Set 3	Safety level	Korea Cosmetic Industry Institute

The standardization of the component name was the most important work in the database construction process, and it was conducted by comparing Product Ingredient data (Set1) and Product Ingredient Standardized (Set2). The standardization process was carried out in three stages. In the first step, if the specific component name in Set1 exists in the standard name or old name record of Set2, it is replaced with the standard name, otherwise, it is filled with null value. Through this process, 70% or more components were standardized. In the second step, the English name of the component name treated as null value in the first step is compared with the standard English name or old English name record of Set2, then the standard name is returned if two words are matched, otherwise null value is still outputted. Up to this stage, more than 95% of the ingredients were standardized. In the last step, the null value of less than 5% was regarded as a non-standardized component and replaced with the original component name in Set1.

3 Conclusion and Future Research

This study has highlighted the problem of the shortage of the information providing system that meets the demand for safety information on baby cosmetics. In order to solve this problem, database of a safety information of baby cosmetics was constructed by integrating the three data: highly reliable ingredient safety level data, all cosmetic ingredient name data, and Product Ingredient Standardized data. In addition, the database was visualized as a web page so that consumers can easily obtain relevant information anytime and anywhere, and it helped them to understand more intuitively by converting and summarizing the safety level. This kind of consumer-oriented technology information and system development is expected to help not only consumers but also future research.

References

1. Sun-Hee Moon et al.: Cosmetic Safety Awareness: A Comparison between General Consumers and Skin Care Workers. Journal of Investigative Cosmetology. 14, 343-350 (2018)
2. Song-I Park, Mee-Ok Choi: Awareness of the Full Ingredients Labeling System of Cosmetics According to the Degree of Interest in Cosmetics: case in the female college in Kwangju. Journal of Korean Beauty Society. 22, 1188-1195 (2016)
3. Hye-Gyoung Koo: The effectiveness of cosmetic ingredients labeling as a method of providing technical information. journal of consumer policy studies. 42, 219-248 (2012)
4. Hae Lee, Ju-Duck Kim: Influence of Internet Reviews about Cosmetics Distributed Online on Consumer Purchase Behavior. 5, 209-218 (2015)
5. Su-Hyun Yoon: Impact of the level of awareness about cosmetic ingredients on purchase behavior. Master’s thesis, Department of Cosmetic & Beauty, Sookmyung Women's University (2015)

A Study on Correlations between Inauthenticity/Job Burnout and Emotional Labor Based on Big Data Analysis: Focusing on Vietnamese Workers

Thi-Hong Nguyen*
*Coordinator of VK Technological Exchange Seminar at Thanh Hoa, Thanh Hoa, Socialist Republic of Vietnam
*Corresponding Author Email: thuhong2903@hotmail.com

Abstract. This research deals with the elements affecting emotional labor and job burnout as well as the influence of inauthenticity over worker's emotional status while performing his/her task. Through conceptual research utilizing the existing literature and questionnaire, empirical research was conducted and the factors such as surface or deep acting, frequency, intensity, duration, and variety of emotional labor were considered as an independent variable associated with the dependent variables inauthenticity and job burnout. At the same time, the demographic variable was defined as a sole variable for the study. The research result has revealed that both actings types indeed had a direct and proportional relationship with the dependent variables isolating the worker from his true feelings. Also, the intensity and the variety of emotional labor has a positive relationship with the same dependent variables whereas its frequency and duration did not. This means that proper supervision or human resource management plan should be established primarily focusing on workers' emotional care to allow them to balance out both actings to the extent that they can perform their tasks without experiencing overwhelming self-inauthenticity or job burnout. As such, this research attempts to find a solution which could ease the intensity of emotional labor along with an effective means of reducing its variety, paying more attention to the psychological state of workers.

Keywords: Emotional Labor, Job Burnout, Big Data; Big Data Analysis; Vietnamese Workers

1. Introduction

Today, the service personnel in every industry are often forced into an environment where they are required to express their feelings appropriately according to the organization's 'rules for emotional expression' while interacting with customers regardless of their own emotional feelings [1-5].

Currently, service industry occupies a large part in the entire industry and is growing continuously and since it is being recognized that customer satisfaction and service are the determinants of a firm's competitiveness, the all-out productization of human has started to progress (Hochschild, 1983). The emotional labor, or rather an appropriate expression of emotions from a service organization's point of view, increases the chance of customers' additional or repeat orders in addition to attracting more customers (Rafaeli & Sutton, 1987; Rafaeli, 1988). For this reason, firms started to intervene in the emotional expression of their employees deeply [6-8].

Emotional labor is 'management of feelings to produce a facial or a body expression which can be observed externally' (Hochschild, 1983), which means that the emotional labor of service personnel is essential for the smooth performance of a service process [6].

The products of both physical and mental labors are goods or new ideas, etc., whereas the products of emotional labor are workers' smile, emotions or feelings.

2. Big Data-Based Empirical Study

The big data analysis can visualize a form by collected unstructured data fragments as a puzzle generates a picture by matching scattered pieces [5]. Accordingly, the market for the big data is becoming larger over time and the data is being used in different areas of our daily lives and much information is shared by the general population. However, since the analysis of big data is very complicated and difficult that sometimes it is quite hard to recognize its meaning and direction, the visualization of big data has come into the picture. Recently, the big data analysis is shifting from AMOS/Python to R/Hadoop [9-13]. The data processing in the Socialist Republic of Vietnam (Vietnam, hereunder) is in an early stage and a variety of problems are needed to be solved [14-15].

The objective of this study goal has been achieved by performing both conceptual and empirical research in parallel, focusing on the latter mainly. The relationships between emotional labor and self-inauthenticity or job burnout were investigated for conceptual study by examining a number of preceding studies and related literature to derive hypotheses whereas an empirical analysis was conducted as an empirical study with a questionnaire based on the conceptual study targeting the workers in Vietnam.

[Table. 1] shows the analysis result of the correlations with individual variables that have gained validity through factor analysis, which was then used to verify some of the causal relationships established in this study.

[Table. 1] The result of correlation analysis

Variable	Mean	Std. Dev.	1	2	3	4	5	6	7	8	9
1. Surface Acting	3.7	0.7									
2. Deep Acting	3.6	0.6	.286**								
3. Frequency1	57.9	30.8	0.105	0.12							
4. Duration1	13	7.9	-0.024	0.151	-.205*						
5. Duration2	6.6	4.8	0.044	0.088	-.328**	.736**					
6. Frequency2	5.5	4.0	0.136	0.096	0.052	-0.098	-0.116				
7. Intensity	3.6	0.6	.387**	.364**	0.088	0.102	0.104	.188*			
8. Variety	3.6	0.8	.206*	.376**	.287**	0.151	0.043	-.206*	.242**		
9. Self-	3.6	0.6	.606**	.610**	.206**	0.106	0.017	0.001	.606**	.455**	

Inauthenticity											
10.Job Burnout	3.5	0.5	.506**	.529**	0.094	0.043	0.101	0.082	.506**	.332**	.540**

The big data analysis result showed that most of the correlations between emotional labor dimensions and the lower dimensions of inauthenticity or job burnout were significant. Meanwhile, observing the correlations between the dimensions, the variables associated with either surface acting or deep acting had a positive (+) relationship mostly. Such results are almost consistent with the hypotheses established earlier and confirm significant correlations between those variables.

Among the variables of emotional labor, 'intensity' had a positive correlation with surface acting (r=.387, p<.01), deep acting (r=.364, p<.01), self-inauthenticity (r=.606, p<.01), and job burnout (r=.506, p<.01) and 'variety' also has the same correlation with them at the level of r=.206, p<.01, r=.276, p<.01, r=.455. p<.01, r=.332, p<.01, respectively. At the same time, surface acting (deep acting) also had a positive correlation with inauthenticity and job burnout at the level of r=.506 p<.01 (r=.610 p<.01) and r=.606, p<.01 (r=.529, p<.01), respectively.

3. Conclusion

The main objective of this study was to investigate the influence of emotional labor on either inauthenticity or job burnout through an empirical analysis, which was achieved through a conceptual study based on the related literature and an empirical study with a questionnaire. The verification of research hypotheses was performed by targeting 155 employees working as a cabin attendant, nurse, and sales personnel. The summary of the empirical analysis result is as follows: first, both surface acting and deep acting had a positive influence on inauthenticity, being consistent with the hypothesis. Second, 'frequency' and 'duration' neither affected inauthenticity nor conformed to the hypothesis. Meanwhile, 'intensity' and 'variety' had a positive effect on inauthenticity and supported the hypothesis. Third, both surface acting and deep acting had a positive influence on job burnout, supporting the hypothesis as well. Finally, while 'intensity' and 'variety' did have a positive influence on job burnout, 'frequency' and 'duration' did not, and the hypothesis was considered to be incorrect.

This study attempted to present some of the important implications drawn from the result of an empirical analysis which focused on the effects of emotional labor on inauthenticity or job burnout.

References

[1] Thi-Hong Nguyen, The Effects of Emotional Labor on Inauthenticity and Job Burnout: Big Data-Based Empirical Study in Vietnam, Proceedings of MITA 2019, National University, Ho Chi Minh City, Socialist Republic of Vietnam, ISSN 1975-4736, 2019, pp.1-2

[2] Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. Science, 349, 255–260, 2015.

[3] Le Cun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature, 521, 436–444, 2015.

[4] Jindal, N.; Liu, B. Identifying Comparative Sentences in Text Documents. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association or Computing Machinery (ACM), Seattle, WA, USA, 6–11 August, pp. 244–251, 2006.

[5] Huh, J. H.; Big data analysis for personalized health activities: Machine learning processing for automatic keyword extraction approach. Symmetry, 10(4), 93, 2018.

[6] Hochschild, Arlie R. ;The managed heart. Berkeley, 1983.

[7] Sutton, R. I., & Rafaeli, A.; Characteristics of work stations as potential occupational stressors. Academy of Management Journal, 30.2, 260-276, 1987.

[8] Sutton, R. I., Rafaeli, A.; Untangling the relationship between displayed emotions and organizational sales: The case of convenience stores. Academy of Management journal, 31, 3, 461-487, 1988.

[9] H. Jung, S. Kim, J.M. Gil, U.M. Kim,; "Processing Continuous Range Queries with Non-spatial Selections," Mobile, Ubiquitous, and Intelligent Computing. Springer, 2014. 31-38.

[10] Kwanho In, Seongkyu Kim, Ung-Mo Kim.; "DSPI: An Efficient Index for Processing Range Queries on Wireless Broadcast Stream," Mobile, Ubiquitous, and Intelligent Computing, Springer, 2014, 39-46.

[11] Man-Kyu Huh, Hong-Wook Huh.; "Genetic diversity and population structure of wild lentil tare," Crop Science, 41.6 (2001): 1940-1946.

[12] Se-Hoon Jung, Jun-Ho Huh.; "A Novel on Transmission Line Tower Big Data Analysis Model Using Altered K-means and ADQL," Sustainability, MDPI, 2019, 11(13), 1-25.

[13] Kim, H. K., So, W. H., Je, S. M.; "A big data framework for network security of small and medium enterprises for future computing," The Journal of Supercomputing, Springer, 2019, 75(6), 3334-3367.

[14] Ngu, Huynh Cong Viet, Jun-Ho Huh.; "B+-tree construction on massive data with Hadoop," Cluster computing, Springer, 2017, 1-11.

[15] Huynh,C.V.N., Kim,J., Huh,J.H.; "Improving the B+-tree construction for transaction log data in bank system using Hadoop," In International Conference on Information Science and Applications, Springer, Singapore, 2017, pp. 519-525.

A Blood Cold Chain System using Blockchain Technologies

Seungeun Kim[1] and Joohyung Kim[1] and Dongsoo Kim[2]

[1] Department of IT Distribution and Logistics
[2]Department of Industrial and Information Systems Engineering
Soongsil University
369 Sangdo-Ro, Dongjak-Gu, 06978 Seoul, Korea
sungkim@soongsil.ac.kr, kjh512@ssu.ac.kr, Corresponding author: dskim@ssu.ac.kr

Abstract. Existing centralized blood management systems have limitations in that they lack detailed blood information and information is not reflected in real time. So, this study designs and implements a new blood cold chain system using blockchain technologies. The proposed system aims to increase the information visibility of blood cold chain system by recording overall information of blood supply and detailed blood information such as blood consumption and disposal to the distributed ledger. In addition, this study proposes directly blood transaction between medical institutions can be performed in case of emergencies. The private blockchain techniques with limited participants are relatively fast and reliable, making it suitable for B2B transactions. Therefore, this study is based on the architecture of Hyperledger Fabric which is one of the private blockchain technologies and implemented by Hyperledger Composer tool.

Keywords: Blood cold chain, Blockchain, Hyperledger, Information visibility

1 Introduction

Blood for blood transfusion directly linked to life depends on the amount of blood donated. Most of the blood is responsible from the government because of the specificity of management. In Korea, the Korean Red Cross Blood Service Headquarters established in 1958 is in charge [1]. Fifteen blood banks and three blood inspection centers are operated, and the blood banks collect, store, and distribute the blood. Two major issues in the domestic blood cold chain are the lack of information visibility in the blood information system and the delay in blood transit time during emergencies due to the centralized blood supply system.

Since blood is highly uncertain, supply agility among them is required to respond quickly [2]. Gaining visibility through information sharing in the supply chain allows for a flexible and agile supply chain. Accordingly, when information visibility on the blood cold chain is secured, real-time information sharing among the participants and transparent information management enable efficient and agile response to demand. Due to the aging population and low birth rate in Korea, the number of people who can donate will decrease, and the number of people who need transfusions will increase. The optimal blood reserve is more than five days a day, but the actual amount is not enough. In addition, in certain areas, the supply range of the blood supply source in charge is widened, and the supply of the blood supply to the medical institution is more than an hour [3]. Because it is very lethal to the patient's life, it is necessary to construct a supply system to cope with urgent demand.

This study presents an architecture that can compare a new blood cold chain based on existing centralized blood cold chain and blockchain technology. This paper describes the design method of blood cold chain system based on the proposed private blockchain technology and proposes the blood transaction between medical institutions so that the medical institution which is distant from the blood bank in the emergency can receive the blood quickly. It is realized by using Hyperledger Composer that information visibility can be ensured by generating and tracking single records for each supply point, and it is confirmed that blood supply time can be reduced in emergency situations.

The rest of the paper is organized as follows. Section 2 reviews previous researches of blockchain and blood cold chain system. Section 3 presents the design of new blood cold chain system and Section 4 shows scenarios based on new blood cold chain system Finally, Section 5 offers conclusions.

2 Related Work

The block chain technology that emerged with the emergence of bitcoin is decentralized distributed database technology. It is possible to share data of each system database on the network and share it in real time in a distributed environment, and it is possible to maintain consistent data by agreement algorithm [4]. A block chain, also called distributed branching technology, consists of a number of blocks, each of which contains several pieces of transaction information. Participants (nodes) of a block-chain network can directly deal with a relay or a third party without verifying the transferred block and adding only approved blocks. Typical block chain-based technologies include Bitcoin Core, Ethereum, and Hyperledger Fabric. Hyperledger Fabric provides three representative services, Membership, Agreement Algorithm, and Chain code. First, the membership service manages the information of the participants to give limited authority to the blockchain network. To do this, we issue participant certificates, transaction certificates, and communication encryption certificates. Second, the Hyperledger Fabric has a different consistency approach than Bitcoin Core or Ethereum. To obtain participants' consensus when updating the data, a consensus algorithm called Practical Byzantine Fault Tolerance (PBFT) is used. In the PBFT scheme, a node plays a role of a leader, sends a request to all nodes including itself, aggregates the response, and confirms the block. Third, chain code is a program for processing transaction execution, and implements processing such as Init (Initialization), Invoke (Transaction execution), Query.

Unlike the 0.6 version of the existing Hyperledger Fabric described above, the largest change in 1.0 of the Hyperledger Fabric used in this paper is the role of the peer [5]. In version 1.0, Validation Peer, which played a major role in consensus algorithm and ledger management, is divided into Endorsing Peer, Committer Peer, and Ordering Noder. The assurance peer performs the task of verifying the transaction proposal, and when the ordering service sequentially arranges for each transaction to store a consistent ledger for the transaction, the committer peer performs the task of storing the ledger through transaction validation.

In the medical field, domestic research using blockchain technology has realized the ease of accessing personal information of patients, and the repository of health checkup results by using Hyperledger Fabric technology for security management, which is characteristic of medical information [6]. In previous study, a new blood supply process was designed using Hyperledger Fabric technology, however it was implemented through transaction execution in this study [7]. In addition, Yang et al. Proposed a study using a block chain for a healthcare system that enhances the security of identity authentication and facilitates access to electronic medical record (EMR) information using smart contract characteristics [8]. Nagurney et al. Presented a new blood supply chain network that minimizes costs and risks by displaying the corruption characteristics of the blood as arc multipliers [9]. Jabbarzadeh et al. Presented a study that designed a network using an optimization model for blood supply in the event of a disaster [10]. In order to preserve and exchange medical records, blockchain technology applications in the medical field are increasingly being implemented by the state and several private companies [11].

3 Design of Blood Cold Chain System

The composition of the blood cold chain system corresponds to the blood management headquarters, the blood bank, the blood center, the transportation vehicle, and the medical institution. Because participants are already defined, the private blockchain approach is appropriate. Fig. 1 shows the existing blood cold chain and the architecture of the proposed new system. The left side of Fig. 1 is the conventional centralized form and the right side of Fig. 1 is the blood cold chain system with the shared distributed branching technology.

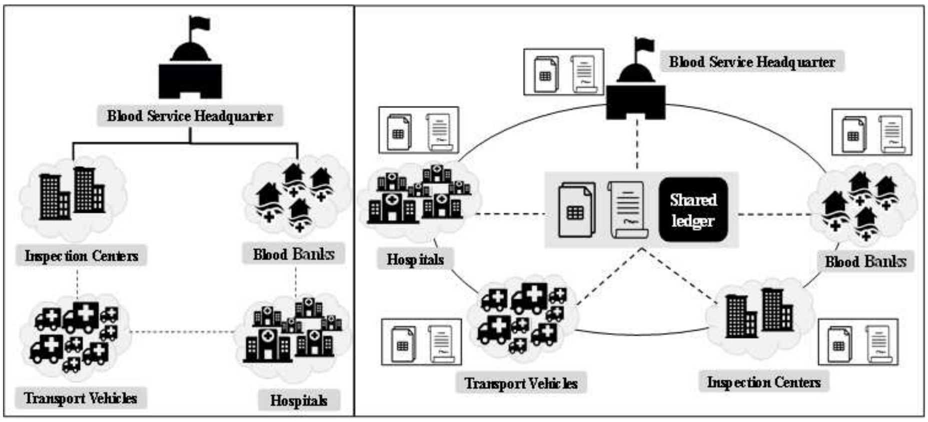


Fig. 1. Conceptual architecture of blood cold chain system

By the blockchain technology, the network participants share the same distributed ledger, and the blood management headquarters plays a role of information inquiry and transaction monitoring, not direct participation in the supply.

The distributed rootstock architecture of the blood cold chain system is shown in Fig. 2. This is based on the block structure of the Hyperledger Fabric, which shows the transactions that are generated each time the blood moves along the base. There are a total of 15 transactions in the Block 1000 that produce blood for blood transfusion, and the transactions that move from the blood bank to the inspection center via the transport vehicle are in Block 1001 and Block 1002. The transaction in which the blood which has been judged as conforming is collected again into the blood bank is that Block 1003 and Block 1004, and Block 1005 and Block 1006 are transferred from the blood bank to the medical institution by the blood order of the medical institution. Finally, Block 1007 shows the use of blood stored in a medical institution.

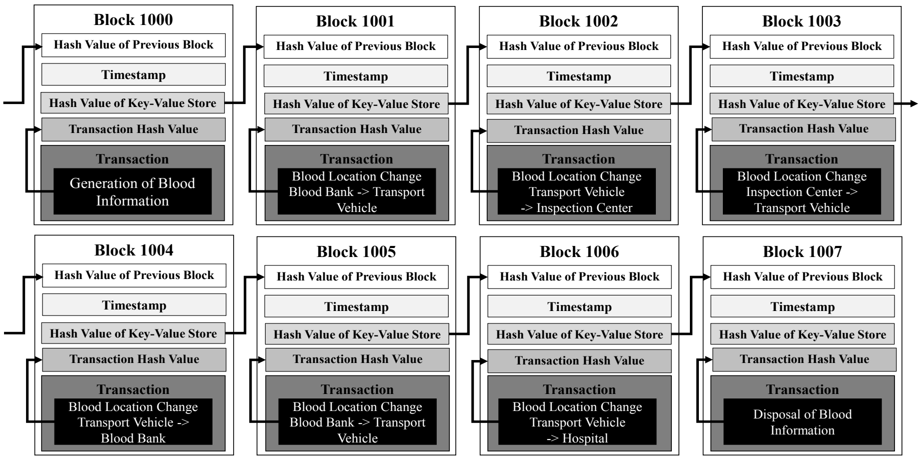

Block 1000
Hash Value of Previous Block
Timestamp
Hash Value of Key-Value Store
Transaction Hash Value
Transaction
Generation of Blood Information
Block 1001
Hash Value of Previous Block
Timestamp
Hash Value of Key-Value Store
Transaction Hash Value
Transaction
Blood Location Change Blood Bank -> Transport Vehicle
Block 1002
Hash Value of Previous Block
Timestamp
Hash Value of Key-Value Store
Transaction Hash Value
Transaction
Blood Location Change Transport Vehicle -> Inspection Center
Block 1003
Hash Value of Previous Block
Timestamp
Hash Value of Key-Value Store
Transaction Hash Value
Transaction
Blood Location Change Inspection Center -> Transport Vehicle
Block 1004
Hash Value of Previous Block
Timestamp
Hash Value of Key-Value Store
Transaction Hash Value
Transaction
Blood Location Change Transport Vehicle -> Blood Bank
Block 1005
Hash Value of Previous Block
Timestamp
Hash Value of Key-Value Store
Transaction Hash Value
Transaction
Blood Location Change Blood Bank -> Transport Vehicle
Block 1006
Hash Value of Previous Block
Timestamp
Hash Value of Key-Value Store
Transaction Hash Value
Transaction
Blood Location Change Transport Vehicle -> Hospital
Block 1007
Hash Value of Previous Block
Timestamp
Hash Value of Key-Value Store
Transaction Hash Value
Transaction
Disposal of Blood Information

Fig. 2. Distributed ledger architecture

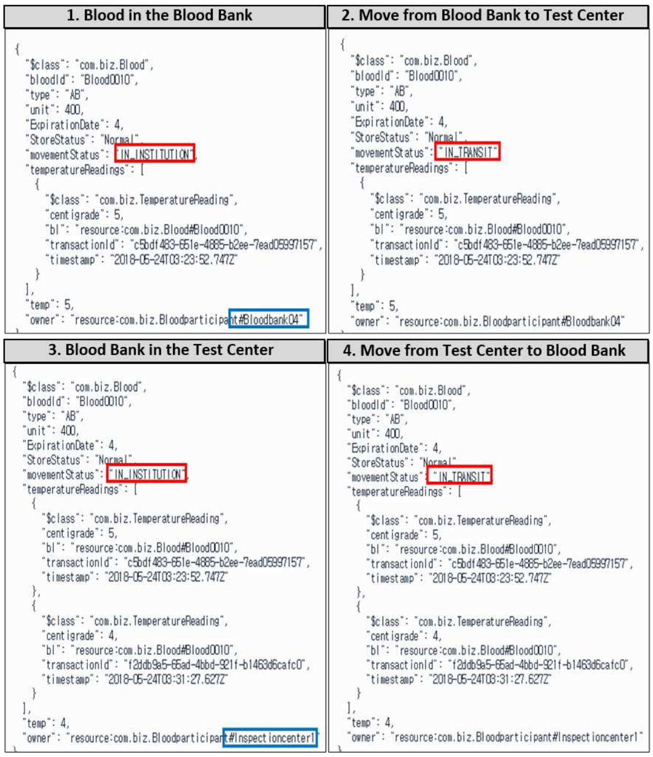
The transactions that are invoked as smart contracts are recorded as logs. The Key Value Store(KVS) in each block performs the task of keeping the transaction processing results up to date. KVS has the same contents in all verification nodes, and the corresponding hash value is recorded in the blockchain. Blood is sensitive to temperature as resources that should be dedicated refrigerator to maintain a constant temperature. Collected blood should be stored at 1 to 6 degrees Celsius, which means that after performing a temperature confirmation transaction, blood that deviates from the storage reference temperature must be discarded. The system attempts to add a temperature confirmation transaction to the blood information so as to determine the blood having a problem in the temperature management as the object to be discarded. When the temperature information is confirmed at a certain point in time, it can be recorded in a transaction through a chain code, which is recorded in the blood property information. Chain code is a program for transaction execution and is based on smart contract execution.

On the other hand, blood may be discarded for reasons such as turbidity and discoloration in addition to the storage temperature. In order to manage it, this Study defined the attribute that indicates the state of preservation in blood information. This attribute is required and must be entered when all initial blood assets are created. In addition, the blood cold chain participant updates the cause if the current blood storage state changes. At this time, if the modified storage state is not normal, the blood is removed through the blood waste transaction by the storage condition.

Blood has a characteristic that the preservation period varies depending on the ingredient preparation. Therefore, in addition to the temperature and the storage condition, the storage period should be continuously checked and the storage period should be discarded. After one day in the system, the total duration of blood storage will be reduced by one day. Like temperature and storage state, the transaction to be discarded is designed when the retention period reaches 0 days.

4 Scenario

This system consists of two scenarios. In the first scenario, the general situation is that the blood bank of the Korean Red Cross Blood Management Headquarters produces blood for blood transfusion, sends blood to the inspection center and checks the stability and supplies blood at the request of hospitals. Movement status was defined to indicate whether blood is currently moving or in the organ.



1. Blood in the Blood Bank

```
{
  "$class": "com.biz.Blood",
  "bloodId": "Blood0010",
  "type": "AB",
  "unit": 400,
  "ExpirationDate": 4,
  "StoreStatus": "Normal",
  "movementStatus": "IN_INSTITUTION",
  "temperatureReadings": [
    {
      "$class": "com.biz.TemperatureReading",
      "centigrade": 5,
      "bl": "resource:com.biz.Blood#Blood0010",
      "transactionId": "c5bdf483-651e-4885-b2ee-7ead05997157",
      "timestamp": "2018-05-24T03:23:52.747Z"
    }
  ],
  "temp": 5,
  "owner": "resource:com.biz.Bloodparticipant#Bloodbank04"
```

2. Move from Blood Bank to Test Center

```
{
  "$class": "com.biz.Blood",
  "bloodId": "Blood0010",
  "type": "AB",
  "unit": 400,
  "ExpirationDate": 4,
  "StoreStatus": "Normal",
  "movementStatus": "IN_TRANSIT",
  "temperatureReadings": [
    {
      "$class": "com.biz.TemperatureReading",
      "centigrade": 5,
      "bl": "resource:com.biz.Blood#Blood0010",
      "transactionId": "c5bdf483-651e-4885-b2ee-7ead05997157",
      "timestamp": "2018-05-24T03:23:52.747Z"
    }
  ],
  "temp": 5,
  "owner": "resource:com.biz.Bloodparticipant#Bloodbank04"
```

3. Blood Bank in the Test Center

```
{
  "$class": "com.biz.Blood",
  "bloodId": "Blood0010",
  "type": "AB",
  "unit": 400,
  "ExpirationDate": 4,
  "StoreStatus": "Normal",
  "movementStatus": "IN_INSTITUTION",
  "temperatureReadings": [
    {
      "$class": "com.biz.TemperatureReading",
      "centigrade": 5,
      "bl": "resource:com.biz.Blood#Blood0010",
      "transactionId": "c5bdf483-651e-4885-b2ee-7ead05997157",
      "timestamp": "2018-05-24T03:23:52.747Z"
    },
    {
      "$class": "com.biz.TemperatureReading",
      "centigrade": 4,
      "bl": "resource:com.biz.Blood#Blood0010",
      "transactionId": "f2ddb9a5-65ad-4bbd-921f-b1463d6cafc0",
      "timestamp": "2018-05-24T03:31:27.627Z"
    }
  ],
  "temp": 4,
  "owner": "resource:com.biz.Bloodparticipant#Inspectioncenter1"
```

4. Move from Test Center to Blood Bank

```
{
  "$class": "com.biz.Blood",
  "bloodId": "Blood0010",
  "type": "AB",
  "unit": 400,
  "ExpirationDate": 4,
  "StoreStatus": "Normal",
  "movementStatus": "IN_TRANSIT",
  "temperatureReadings": [
    {
      "$class": "com.biz.TemperatureReading",
      "centigrade": 5,
      "bl": "resource:com.biz.Blood#Blood0010",
      "transactionId": "c5bdf483-651e-4885-b2ee-7ead05997157",
      "timestamp": "2018-05-24T03:23:52.747Z"
    },
    {
      "$class": "com.biz.TemperatureReading",
      "centigrade": 4,
      "bl": "resource:com.biz.Blood#Blood0010",
      "transactionId": "f2ddb9a5-65ad-4bbd-921f-b1463d6cafc0",
      "timestamp": "2018-05-24T03:31:27.627Z"
    }
  ],
  "temp": 4,
  "owner": "resource:com.biz.Bloodparticipant#Inspectioncenter1"
```

Fig. 3. Blood movement status and proprietary information changed by blood movement

In addition, ownership information of the blood is defined, and the first owner of the blood corresponds to the blood bank. In Fig. 3, the first data represents the first blood, indicating that the movement is within the organ and the owner is the blood bank. The second data shows that when the blood goes to the inspection center, it changes to the status of moving. After moving from the third data to the inspection center, the owner can confirm that the inspection center has been changed. When a transaction is performed to check the blood temperature within the inspection center, the results are recorded in the blood asset details. The fourth data shows that after being judged to be suitable for blood for blood donation, it moves again to the blood bank.

In the second scenario, In the event of an emergency, certain hospitals are out of stock and need to supply blood within 60 minutes. Assuming that certain conditions are met, blood is ordered and supplied to nearby hospitals. Under normal circumstances, hospitals receive blood from the blood bank. However, hospitals that are far away from the blood bank in emergency situations have longer delivery times. To this end, business logic is defined in this system that enables the transaction of blood between hospitals under the circumstances. The logic can be divided into two parts. First, the requesting hospital should be in a state of emergency. Second, the requested hospital should be in a state where surplus blood exists to supply blood to the neighboring hospitals(Hospital B). At this time, whether an emergency occurs or not is defined as a property of the participant registry, and the hospital can modify this information at any time. If both logic is not established at the same time, the system notifies the user that the transaction cannot be concluded.

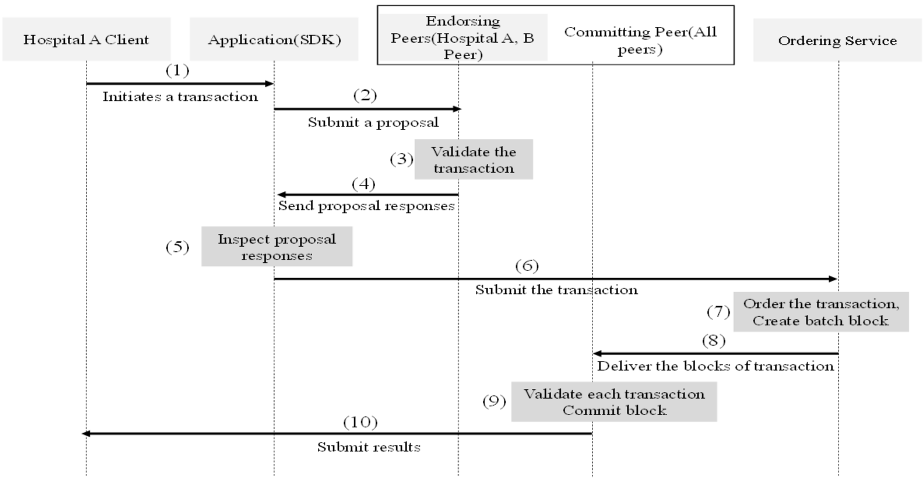


Fig. 4. Screen of successful inter-hospital blood request transaction

The role of verifying the requested transaction is the peer of the hospital A and the neighboring hospitals(Hospital B) of the hospital A, because the Hyperledger Fabric uses the deconvolution algorithm by the specific verification node. In Fig. 4, (1) shows that the client of hospital A requests a transaction that blood is needed. (2) and (3) show that when the proposal is delivered to the assurance peer, the peers of hospital A and the neighboring hospitals(Hospital B) corresponding to the assurance peer execute the chain code to verify the transaction. In (4) and (5), the application examines the results received from the guaranteed peer. (6) shows that the application submits the transaction to the ordering service. In (7) and (8), the ordering service defines the transaction sequence, creates one block and sends it to all peers. (9) shows that each peer verifies the transaction and confirms the transaction after verifying that the guarantee condition is met. (10) shows that the neighboring hospitals(Hospital B) capable of ultimately supplying the surplus blood supply blood to the hospital A.

5 Conclusion

In this study, we designed and implemented a blood cold chain system based on private blockchain technology to achieve two goals: securing information visibility and reducing blood supply time. First, recording and sharing the information of each time the blood is moved and the information that is consumed and discarded to the dispersal director in real time enables the efficient management of the current blood business by tracking the number of units from the supply to the final blood transfusion do. This system, which prevents information forgery and tampering, can solve the problems that may arise due to relying on medical institutions and medical staff to evaluate the appropriateness of blood use, and thus it will be more transparent to operate the blood business as a humanitarian resource. Second, a system that supports blood transfusion between medical institutions through an agreement process for urgent demand can reduce the blood supply time which is directly connected with the patient's life. When hospitals are overstocking their blood type stocks, most of the time they are holding the blood and reducing the blood order. Therefor they can use the surplus blood efficiently.

Future studies will first examine whether the proposed design and implementation can be interfaced with actual blood information systems, taking into account the cost and network configuration. Second, additional information, block size, and possible hacking need to be taken into account in the actual hospital-to-hospital blood transaction. In order to build this system, participation in institutions such as blood management headquarters, blood bank, blood inspection center, and hospital is required. We will further expand the scenarios presented in future studies that take into account national policy and operational guidelines for blood operations to increase their applicability in the medical field.

Acknowledgments. This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1D1A1B05029080).

References

1. Korean Red Cross Blood Service Headquarters, http://www.bloodinfo.net/main.do.
2. Hess, J. R., Thomas, M. J. G.: Blood use in war and disaster: lessons from the past century. Transfusion, vol. 43, 1622--1633 (2003)

3. Kim, S.: Agile Blood Supply Chain Design Considering Golden Time. Korean Journal of Logistics. vol. 24, 4, 27--40 (2016)
4. Crosby, M., Pattanayak, P., Verma, S., Kalyanaraman, V.: Blockchain technology: Beyond bitcoin. Applied Innovation. 2, 6--10 (2016)
5. Yoshiharu. A., et al.: Blockchain Structure and Theory. (2017)
6. Kim, M.: Blockchain based Health Checkup Results CDA Viewing Service. Kyungpook National University, Daegu, Korea (2017)
7. Kim, S., Kim, D.: Design of an Innovative Blood Cold Chain Management System Using Blockchain Technologies: ICIC Express Letters. Part B: Applications. vol. 9:10, pp. 1067--1073 (2018)
8. Yang, S.: Proposal for Smart Contract method for domestic medical system based on the colored coin. Soonchunhyang University, Asan, Korea (2017)
9. Nagurney, A., Masoumi, A. H., Yu, M.: Supply chain network operations management of a blood banking system with cost and risk minimization. Computational Management Science, 9.2, pp. 205--231 (2012)
10. Jabbarzadeh, A., Fahimnia, B., Seuring, S.: Dynamic supply chain network design for the supply of blood in disasters: a robust model with real world application. Transportation Research Part E: Logistics and Transportation Review, 70, pp. 225--244 (2014)
11. Choi, H.: Use of block-chain technology in the healthcare industry, Korea Health Industry Development Institution (2017)

Fuzzy Logic-based Value of Information Assessment to Support Effective Decision Making in Battlespace[1]

Kunyoung Kim[1], Mye Sohn[1]*, and Gyudong Park[2]

[1] Department of Industrial Engineering, Sungkyunkwan University
Suwon, Korea
{kimkun0,myesohn}@skku.edu

[2] 2nd R&D Institute, Agency for Defense Development
Seoul, Korea
{iobject@add.re.kr}

Abstract. To support efficient and effective decision-making of the commanders, it is required to deliver the appropriate information in a timely manner. However, they may suffer from receiving the huge amount of information to determine whether they are necessary for the decision-making. To alleviate the burden, we propose a fuzzy logic-based Value of Information (VoI) assessment method. In addition, we devise the assessment metrics of the VoI. To illustrate the background of the derivation of the assessment metrics, we develop an illustrative scenario about surveillance situation of the coastal area.

Keywords: Value of information, Fuzzy logic, Situational awareness, Decision making

1 Introduction

In the information age, information superiority became a critical factor to dominate adversaries in warfare [1]. To maintain information superiority in the battlespace, it is required that heterogeneous information from various sources should be integrated automatically and delivered to the commanders and the staffs in a timely manner [2]. However, the commanders and their staffs may perform different missions and roles depending on the operational situation. Without of this in mind, disseminating all information to all stakeholders can hinder efficient and effective decision-making. To resolve the problem, researchers have introduced the concept of Value of Information (VoI) that can filter and prioritize the information in the operational situation [3, 4, 5]. However, the research has the following limitations. First, they do not provide specific methods to calculate VoI. The reason for this is that their goal is to provide fundamental definitions and directions for VoI assessment on the battlefield. Second, they cannot provide the precise assessment of VoI since they use standardized metrics

[1] This research was supported by C2 integrating and interfacing technologies laboratory of Agency for Defense Development (UD180014ED).

* Corresponding author

regardless of the type of information. Last but not least, they cannot consider the uncertainty and vagueness of the operational situation.

To overcome the limitations, we propose a fuzzy logic-based VoI assessment method that can reflect characteristics of different types of information and uncertainty and vagueness of the operational situation. First, we identified VoI assessment metrics which can be applied to battlespace environment. Second, we developed fuzzy logic-based methods to calculate VoI metrics.

The paper is structured as follows. In section 2, we have developed an illustrative scenario to show the necessity of VoI. Section 3 describes the VoI assessment method by providing metrics and detailed methodology. Finally, Section 4 puts forth the conclusions and suggests further research.

2 Illustrative Scenario

Battalion A, B, and C perform coastal defense of Busan area, which belong to regiment D of Republic of Korea Army. Zone A, B, and C are jurisdictions of Battalion A, B, and C. In the engagement situation, two unknown objects located in zone B are detected by thermal observation devices which are being operated by the battalion B. The overview of the scenario is depicted in Fig. 1.

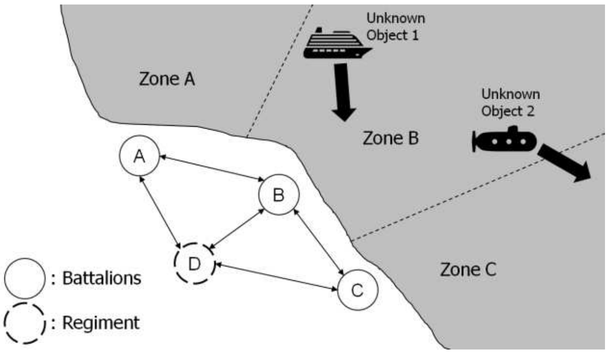


Fig. 1. Overview of illustrative scenario

In this situation, following reports and messages are delivered to the commander of battalion B as Information Objects (IOs).

- **IO 1-1: Track report of unknown object 1** from staffs in battalion B
- **IO 1-2: Track report of unknown object 2** from staffs in battalion B
- **IO 2: Highway situation report** from Ministry of Transport
- **IO 3: Operation order** from regiment D
- **IO 4: Support request** from battalion C

We suppose that each IO follows FM 6-99.2 U.S. Army Report and Message Formats [6]. To deliver these IOs to the commander of battalion B with their priority, VoI of IOs should be calculated.

- **Comparison between IO 1-1 and IO 1-2**

In case of two track reports, unknown object 1 is coming close to battalion B while unknown object 2 is getting out of zone B. That is, IO 1-1 has higher value than IO 1-2 to the commander of battalion B in the perspective of importance of each IO's contents.

- **Value Assessment of IO 1-1**

To assess the impact of the unknown object 1, not only location, speed and heading, which are included in track reports, but also detailed information on the object, such as the type of the ship and mounted weapons, should be considered. At this time, to assess the value of IO 1-1, how much the IO cover the necessary information to understand the situation should be considered.

- **Comparison between IO 1 and IO 2**

Since the mission of battalion B is coastal defense, the track reports (IO 1) have the higher value than the highway situation report (IO 2) to the commander of the battalion B in the perspective of mission relevance. However, if the commander requested information of highway situation, IO 2 has higher value in the perspective of request relevance.

- **Comparison between IO 3 and IO 4**

If the support request (IO 4) is relatively urgent than the operation order (IO 3), IO 4 has higher value than IO 3 in the perspective of timeliness. However, since battalion B is included in regiment D, IO 3 has higher value in the perspective of command structure. Also, these types of IOs should not be accessible to all stakeholders. That is, it is important to deal with the security level of the IOs and access qualifications of the commanders.

3 Value of Information (VoI)

3.1 VoI Definition

The concept of VoI is widely used in the field of decision making for business, and several researchers introduced this concept to military domain. These researchers defined VoI according to their own perspective, but there is no clear one to explain VoI in the battlespace. Therefore, we define VoI as the *relative* value assessed from the relation between an IO and information user in specific context. That is, the VoI of an IO differs from the information user to whom the IO is delivered and the context which the IO is consumed.

3.2 VoI Assessment Metrics

Considering the former studies [3, 4, 5] and Network Centric Operations (NCO) Conceptual Framework [7], we establish seven main metrics to assess VoI of each IO in the battlespace as mentioned as follows:

- **Importance** reflects the intrinsic significance of an IO to each information user. Different features are used to calculate different types of IO. For example, features like ‘geographic proximity’ and ‘approach speed’ can be used to calculate importance of track reports.
- **Completeness** indicates the coverage of the IO for the information user to fully aware the situation. To calculate completeness, both contexts of the information user and contents of IO should be considered.
- **Mission/Role relevance** indicates how much an IO is related to mission or role of the information user. Mission/Role relevance is calculated based on the type of the IO and that of information user. For example, for the commander of the coastal defense battalion, ‘track report’ has higher relevance than ‘highway situation report.’
- **Request relevance** denotes the extent to which an IO can fulfill the information user’s request. When the information user does not request for specific information, VoI is calculated based on the other six metrics.
- **Timeliness** denotes how soon an IO should be delivered to the information user. To calculate timeliness, both contents of an IO and information user context should be considered.
- **Commandability** reflects military hierarchy and command structure. To assess commandability, rank of the information provider and that of the information user should be considered.
- **Confidentiality** reflects security level of the information and access qualification of the information user. That is, it limits the range of information users who can access to certain types of information.

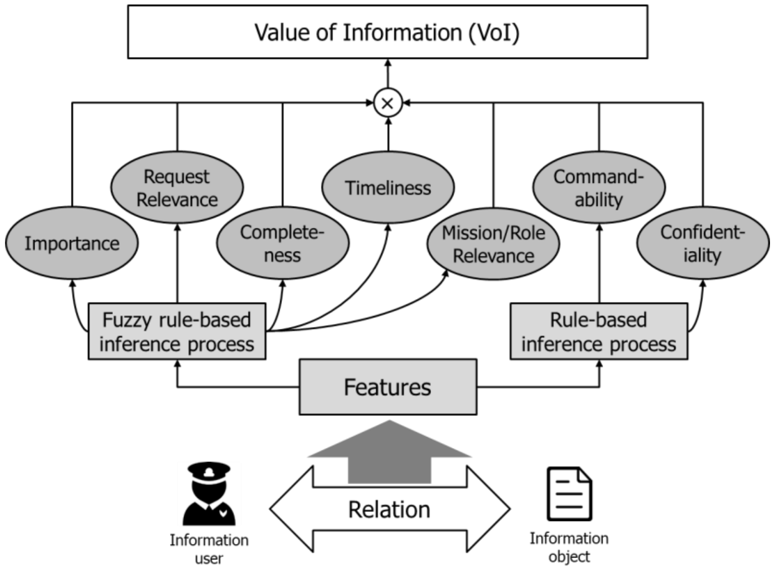


Fig. 2. Assessment process of VoI

3.3 Fuzzy Logic-based VoI Assessment

Decision making in battlespace has the following characteristics. First, information in battlespace may contain uncertainty and incompleteness. Second, decision making is performed mainly based on insight and knowledge of the decision maker. To reflect these characteristics, fuzzy logic can be applied to support decision making in battlespace [8]. Based on seven metrics we have established, we propose fuzzy logic-based VoI Assessment process as depicted in Fig. 2.

Step 1. Calculation of features

Based on the relation between an information user and an IO, features for VoI assessment are calculated. In case of our illustrative scenario, when the commander of battalion B receives track reports of unknown objects, features like 'geographic proximity' and 'approach speed' can be calculated based on contents of the report and context of the commander, such as 'location,' 'speed,' and 'direction.'

Step 2. Assessment of VoI metrics

Using the features we have calculated, assessment process of seven metrics is performed. Since 'commandability' and 'confidentiality' do not contain uncertainty, these metrics can be assessed from rule-based inference process. On the other hand, we introduced fuzzy rule-based inference process to assess the other metrics because these metrics contain uncertainty. To apply fuzzy rules, features are normalized and transformed to fuzzy values. For example, fuzzy values of the geographic proximity are 'Very close,' 'Relatively close,' 'Relatively distant,' and 'Very distant' and their membership functions are shown in Fig. 3. Then fuzzy rules are applied to fuzzy values of each feature, and therefore VoI metrics can be assessed. Finally, VoI can be derived by calculating weighted average of the values of metrics.

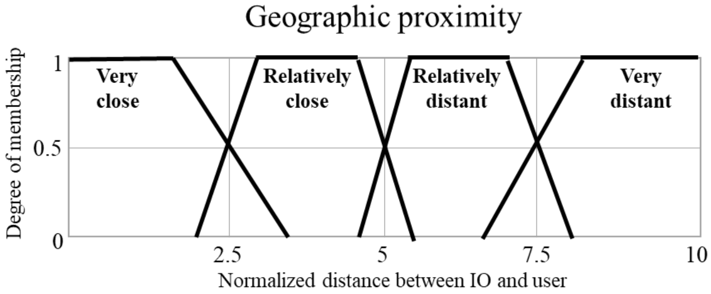


Fig. 3. Example of membership function for geographic proximity

4 Conclusions and Future works

In this paper, we defined VoI in the battlespace and identified main metrics to assess VoI. Also, we proposed fuzzy logic-based VoI assessment process to reflect uncertain and incomplete nature of IOs of battlespace. Using this method, IOs can be delivered to the commander in the precedence determined by VoI. However, there are several limitations. First, it is hard to model implicit knowledge of the domain experts. Second, battlespace contains high level of irregularity. To overcome these limitations, we will develop a prototype by searching relevant literatures and interviewing domain experts, and concretize the method based on actual military data in our future work.

References

1. Alberts, D.S., Garstka, J.J., Stein, F.P.: Network Centric Warfare: Developing and Leveraging Information Superiority. Assistant Secretary of Defense (C3I/Command Control Research Program) Washington DC (2000)
2. Looney, C.G.: Exploring Fusion Architecture for a Common Operational Picture. Information Fusion 2, 251--260 (2001)
3. Bisdikian, C., Kaplan, L.M., & Srivastava, M.B.: On the Quality and Value of Information in Sensor Networks. ACM Transactions on Sensor Networks (TOSN) 9, 48 (2013)
4. Suri, N., Benincasa, G., Lenzi, R., Tortonesi, M., Stefanelli, C., Sadler, L.: Exploring Value of Information-based Approaches to Support Effective Communications in Tactical Networks. IEEE Communications Magazine 53, 39--45 (2015)
5. Michaelis, J.R.: Requirements for Value of Information (VoI) Calculation over Mission Specifications. In: SPIE 10207, Next-Generation Analyst V, 102070M (2017)
6. FM 6-99.2 U.S. Army Report and Message Formats, https://usacac.army.mil/sites/default/files/misc/doctrine/CDG/cdg_resources/manuals/fm/fm6_99x2.pdf
7. Network Centric Operations Conceptual Framework Version 1.0, https://www.hsdl.org/?view&did=446190
8. Hanratty, T.P., Newcomb, E.A., Hammell II, R.J., Richardson, J.T., Mittrick, M.R.: A Fuzzy-based Approach to Support Decision Making in Complex Military Environments. International Journal of Intelligent Information Technologies 12, 1-30 (2016)

Permissioned Blockchain Network and Hyperledger in Manufacturing Industry

Salman Qavi[1], Dilnozkhon Imomalieva[2], Wookey Lee[2]

[1] Department of Computer Science, Stony Brook University.
Stony Brook, NY 11794, USA
[2] Department of Industrial Engineering, Inha University.
100 Inha-ro, Nam-gu, Incheon 22212, Republic of Korea.
sqavi@cs.stonybrook.edu, dina@inha.edu, trinity@inha.ac.kr

Abstract. The need for transparency and traceability is a vital business challenge in manufacturing and maintaining supply chains both locally and globally [1]. Many companies and buyers have little to no information on their second and third tier suppliers and customers. Particularly in the automotive and vehicle engine manufacturing industry, the need for transparency and trust have become a concerning issue. Blockchain as a distributed ledger system can improve transparency and traceability within every tier of the manufacturing supply chain. In this paper, we demonstrate how blockchains can improve transparency and traceability through the implementation of Hyperledger Fabric, a framework that facilitates permissioned and private blockchains, in the production and tracking of car engine blocks by a car manufacturer. We also explore the key challenges and limitations we discovered during the implementation of these blockchain networks. The powerful use of blockchains to track each part of a vehicle through its manufacturing process until reaching the end buyers not only facilitates the buyers and gains their trust, but rather it works both ways as car manufacturers are directly able to know the specific details about the raw materials they are receiving from their parts suppliers and about the specific preferences of customers who are purchasing and using their engine blocks.

Keywords: Permissioned Blockchain Networks, Hyperledger Fabric, Distributed Ledger System, Federated Blockchains, Chain Code, 51% Attack

1 Introduction

The need for transparency and traceability is a vital business challenge in manufacturing and maintaining supply chains both locally and globally [1]. Many companies and buyers have little to no information on their second and third tier suppliers. Incidents in the past decades that illustrated that even tight and expensive security mechanisms are unable to guarantee complete data security, thus leaving organizations at potential risks. The arrival of blockchain comes to the rescue as a blessing. At its core, a blockchain is a decentralized distributed system, which is a collection of autonomous components (computers) that appears to its users to run as a single coherent system as in Fig. 1 [2].

Both blockchain and Hyperledger are emerging concepts and technologies. A blockchain protocol runs on top of the Internet on a peer-to-peer network (i.e. the Internet) of computers (called nodes) that run the protocols individually. Blockchains can be categorized into permissioned and permissionless blockchains, public and private/federated blockchains, and other sub-categories. While Bitcoin is a permissionless and public blockchain - used in the exchange of digital cryptocurrency, there is Hyperledger which is a permissioned and restrictive blockchain but allows private channels to be created for communication among specific participants only.

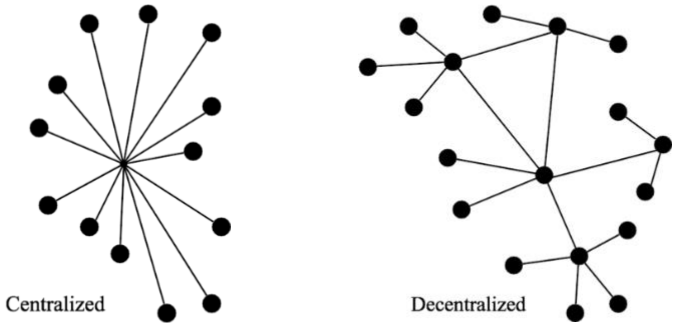


Fig. 1. Diagrammatic representation of (*centralized)* and (*decentralized)* networks. A (*centralized)* network has a single point of failure and resources are shared all the time.

The implementation of distributed ledger technology can improve transparency and traceability issues within every tier of the manufacturing supply chain through the use of immutable records of data or items, distributed storage of the records, and controlled user accesses, and Hyperledger Fabric provides an excellent opportunity as it facilitates permissioned networks [1]. This study exemplifies how distributed ledger technology, such as Hyperledger Fabric, can facilitate the manufacturing industry with an example of blockchain implementation for car engine tracking during car manufacturing process.

2 Experimentation

In order to test the viability of blockchains in the manufacturing industry, we choose a car manufacturer in a developing country[1]. There is a lack of trust in the automotive industry as the distributors and customers have little to no information on the suppliers of raw materials which dictates the quality of parts like tires, engines, etc. and their durability besides many other important factors that decide the price of the cars. Currently, the car company manufactures several models of model X including X-D and X-L[2].

[1] The name of the car manufacturer and the country it is located are kept undisclosed upon request of the car manufacturer.

[2] The original model names are kept undisclosed, instead X, XD, and XL are used to represent the different models the company manufactures.

In Hyperledger Fabric, assets (e.g. engine blocks) that are managed on the blockchain was defined by a model of key-value pairs. The concept of a chain code or smart contract was then implemented based on the business logic on the assets (e.g. engines, tires) and the owners (suppliers of raw materials and the car manufacturer's employees). The chain code can be implemented in high-level programming languages such as Go, Java or Node.js that defines the rights to read and alter any part of the smart contract. It is this place where the information will be stitched during the manufacturing process.

The execution of a chain code function can read and return asset information, create or alter stored information, and store new information in the local ledger database. When all changes were finalized, the changes were proposed to the blockchain network for endorsement and inserted into the blockchain after the endorsement had taken place.

Channels were created to provide privacy. If an entire chain code is deployed on a single channel then all suppliers of raw materials and distributors as well as the end-buyers can see the details of each other's transactions which is not desirable, for instance, in circumstances when the car manufacturer provides discounts to a specific distributor. As a result, several channels were created to accommodate private business transactions between the manufacturer and the different parties - raw materials suppliers and distributors. In Hyperledger Fabric, each participant (i.e. peer node) within a channel keeps a copy of the ledger thus creating a blockchain data structure for the existing channel.

3 Implementation

The scheme of network participants, transactions and events was defined in Hyperledger Composer Modeling Language and flows of each transaction flows were implemented on an API through JavaScript code. In addition, frequently used queries on the stored data were defined in the Composer Query Language, a SQL-like language [3]. All required files were packaged to a Business Network Definition (BND) or .bna file.

The prototype and demo were built on Composer Playground that provides a user-friendly and modern web interface to access configurations of the Composer Command-Line-Interface (CLI). In order to track the engine blocks from their manufacturing to distribution stages, the manufacturers, dealers and customers were added as network participants and engine blocks and vehicles as assets [3].

The raw material suppliers, distributors of cars and the car manufacturer are identified as organizations in the blockchain network. The Hyperledger Fabric chain code that we created to demonstrate the feasibility and usability of our proposal facilitates the following functionalities:

a) the production of a car's engine block with a unique serial number
b) transfer of engine block from the car manufacturer to a dealer after production
c) tracking the car with its unique serial number
d) the installation of an engine
e) block into the registered vehicle that buyers can track.

4 Results

The transparency of the Hyperledger blockchain enabled the suppliers of parts, the car manufacturer as well as distributors and dealers to find out the manufacturing and installation dates, serial number, slot number, location of manufacture, and other specific details of the engines.

From a survey conducted where car manufacturing company employees and top-level executives, car and engine parts suppliers, distributors and end-buyers participated, it was discovered that the implementation of blockchain not only increased transparency and traceability significantly but also the trust of the distributors about the manufacturer and of the manufacturer about the suppliers of the parts as it lessened the risks of counterfeit parts to be used during the manufacturing process of the car engines.

5 Limitations

In spite of all the benefits and potentials in Hyperledger Fabric implementation for car engine tracking in the manufacturing process, several key challenges and limitations were also discovered:

a) Centralization: participants, particularly, distributors expressed their fear that manufacturers can exploit the blockchain networks at their will leading to the centralization of the networks and such networks' vulnerability to 51% attack.

b) Processing power and time: financial information needs to be encrypted before they are stored into the blockchain and decrypted to be read or accessed. Thus, the computing capabilities of the components of the distributed systems need to be interoperable (e.g. participants most likely will need to be running the same version of Operating System and same encryption/decryption algorithms).

c) Storage: blockchain eliminates the need for a centralized server to store all data, but as more and more data are collected and with the emergence of Big Data in the manufacturing industry, a significant portion of the data will need to be stored on the end devices (i.e. Edge/Fog Computing) for efficient searching and processing (insertion, retrieval and modification) performed on the data, but these end devices have only limited storage capabilities.

d) Lack of skills: initially, most people in the survey were not able to understand the concept of blockchains or distributed ledger systems, particularly those who work in non-IT sectors. As a result, it may require more time and effort to teach them the day-to-day operations on the blockchains and to equip them with the skills necessary for secure and efficient updating of the systems when required.

e) Legal and Compliance Issues: the use of blockchains in manufacturing is a completely new field, and there aren't any real standards against which to measure the performance, quality or vulnerability of the blockchains [4]. This can lead to difficult situations when the manufacturers and the participants (e.g. distributors or raw material suppliers) cannot come to an agreement on the legal or compliance issues.

6 Future Work

The resulting blockchain network was only executed locally for one manufacturer; we did not expand the configuration of the peer organizations or the ordering service. In the future, we can expand and execute the network globally, and the performance of the global blockchain network can be evaluated with several manufacturers and many mutual distributors and raw materials suppliers among them. In addition, mechanisms to reduce, if not to eliminate, the possibility of a Majority Attack are to be devised and implemented as other participants within the channels fear the possibility of illegal data manipulation and control over the network or even getting banned from channels if disagreement arises with the manufacturer or the host organization.

7 Conclusion

Through the implemented blockchain application case about the production and tracking of engine blocks, we were able to demonstrate the powerful use of blockchains to track each parts of a car through its manufacturing process until reaching the end buyers. It not only facilitates the buyers and gains their trust, but rather it works both ways as the car manufacturer is now directly able to know who their customers are and their preferences. In addition, we were also able to discover the key challenges and limitations of implementing Hyperledger Fabric in the manufacturing industry.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IICTP) grant funded by the Korea government (MSIT) (2019-0-00136, Development of AI-Convergence Technologies for Smart City Industry Productivity Innovation).

References

1. Abeyratne, S.A., Monfared, R.P.: Blockchain ready manufacturing supply chain using distributed ledger. International Journal of Research in Engineering and Technology, 05(09), 1—10. (2016)
2. Tanenbaum, A.S., Steen, M.van: Distributed Systems: Principles and Paradigms. Pearson Prentice Hall, Amsterdam (2006)
3. Verhoelen, J., Implementing a Blockchain Application with Hyperledger Fabric & Composer, https://blog.codecentric.de/en/2018/04/blockchain-application-fabric-composer
4. Olsen, P., Borit, M., Syed, S.: Applications, Limitations, Costs, and Benefits Related to the Use of Blockchain Technology in the Food Industry. (2019)

Data Visualization for Assessment of Block-Based Programming

Tserenpurev Chuluunsaikhan[1], Aziz Nasridinov[1], Inseong Jeon[2], Ki-Sang Song[2]

[1] Department of Computer Science, Chungbuk National University, Cheongju, Korea
[2]Department of Computer Education, Korea National University of Education, Cheongju, Korea
{teo, aziz}@chungbuk.ac.kr, {jinsung4069, kssong}@knue.ac.kr

Abstract. Entry is one of the block-based programming platforms that is developed in Korea. Teachers can teach code to students while they are playing using the platform. The problem is teachers have to work with many students. In order to increase the time to spend for children, last time we presented Chentry that is an automated evaluation system based on Entry. Chentry uses logs of Entry platform to assess student performance by use of the Levenshtein distance algorithm. Because of the logs, we collect a large amount of data. In order to acquire relevant information from the collected data, we need to analyze the data. This paper introduces a data visualization system based on Chentry's data. The system displays analyzed information in various visualization graphs. Our system helps teachers to improve their teaching strategies.

Keywords: Entry education platform, block-based programming, Automated assignment and assessment, Data visualization

1 Introduction

Big data is not far from us. We all are data sender, also receiver. Human is collecting a large amount of data at every moment in all industries. We increase world data using social networks, online search, shopping, bank transition, learning, etc. In education, Big data is an essential topic too. Based on the big data, researchers work in fields like teaching, assessing, grading, learning, etc. Only having big data cannot solve any problems. We need to acquire relevant information from big data. For that, have to understand big data. That's why data visualization is an integral part of the big data.

Last time, we presented a proposed system named Chentry that is an extension of Entry platform. Entry is an education platform created to help anyone learn to code [10]. Chentry means checking entry. Our system provides convenient views for students and teachers to create assignments and assess students' work. We use logs of Entry platform to assess student's performance by use of the Levenshtein distance algorithm. The system saves every snapshot of students' code. That means we collect a large amount of data. In order to acquire relevant information from the collected data, we need to analyze the data.

This paper introduces a data visualization system for assessment of block-based programming. Data visualization is a graphical representation of data. We are using visual tools like charts, graphs to see and understand trends, outliers, and patterns in data. In that case, we can analyze massive amounts of information and make data-driven decisions.

2 Related Study

[3] introduces CodeMaster, an automatic assessment, and grading system. CodeMaster automatically assesses and grades projects programmed with App Inventor and Snap!. It uses a rubric measuring computational thinking based on static code analysis. The rubric is based on the framework for assessing the development of computational thinking proposed by Brennan & Resnick (2012). Rubrics also provide objective bases for grading by converting rubric scores to grades.

They introduce the basic concept and pipeline of traffic data visualization in [8]. They use the datasets generated and collected by sensors in traffic vehicles or monitors installed along the roads. They used the datasets generated and collected by sensors in traffic vehicles or monitors installed along the roads. The datasets include GPS data of vehicles, GSM locations or cell station records of human mobility, and video/image/counting records of surveillance devices. They provide an overview of relevant visualization techniques and visual analysis system using unstructured data.

In [9], they describe and evaluate the performance of a Big data architecture applied to large-scale knowledge graph visualization. In this work, they address the performance problems of large graph data exploration. They use DrugBank, DBPedia for their experiments. Also, they use Apache Spark and GraphX for visualization.

3 System Design

Fig. 1 shows overall structure of the proposed data visualization system. In this section, we describe our system in detail. Subsection 3.1 explains View Module, Main Module, and Database module. API module is explained in subsection 3.2. Also, subsection 3.3 describe Visualization module.

3.1 Database

First, Chentry system stores introduction information of teachers, students, classes, and login credentials. Also, Chentry collects snapshots from students programming code. The data that is generated from Chentry is stored in MySQL. By writing this study, around 5 teachers, 100 students use this system. In the future, system users will be added. In that case, the volume of data increases quickly.

3.2 Data Processing

Jersey RESTful Web Services framework is open source, production quality, framework for developing RESTful Web Services in Java that provides support for JAX-RS APIs and serves as a JAX-RS (JSR 311 & JSR 339) Reference Implementation [12]. JSON (JavaScript Object Notation) is a popular format for transporting data. We created a rest API using Jersey in this study. By using the rest API, we can use data to many fields from one source easily. The rest API loads data from the database using query and shows processed data in JSON.

3.3 Data Visualization

On the data visualization, we use D3.js, morris js. D3.js is a JavaScript library for manipulating documents based on data. This library drives graph using HTML, CSS, and SVG. Morris is a library that draws graphs like line, bar, area, donut. Comparing to the D3.js, Morris is very simple and easy. Because of it, Morris doesn't have many options. D3.js rawer than Morris. We can create any kind of graphs on D3.js, but that is not that easy to like Morris. In the next section, we describe it in detail.

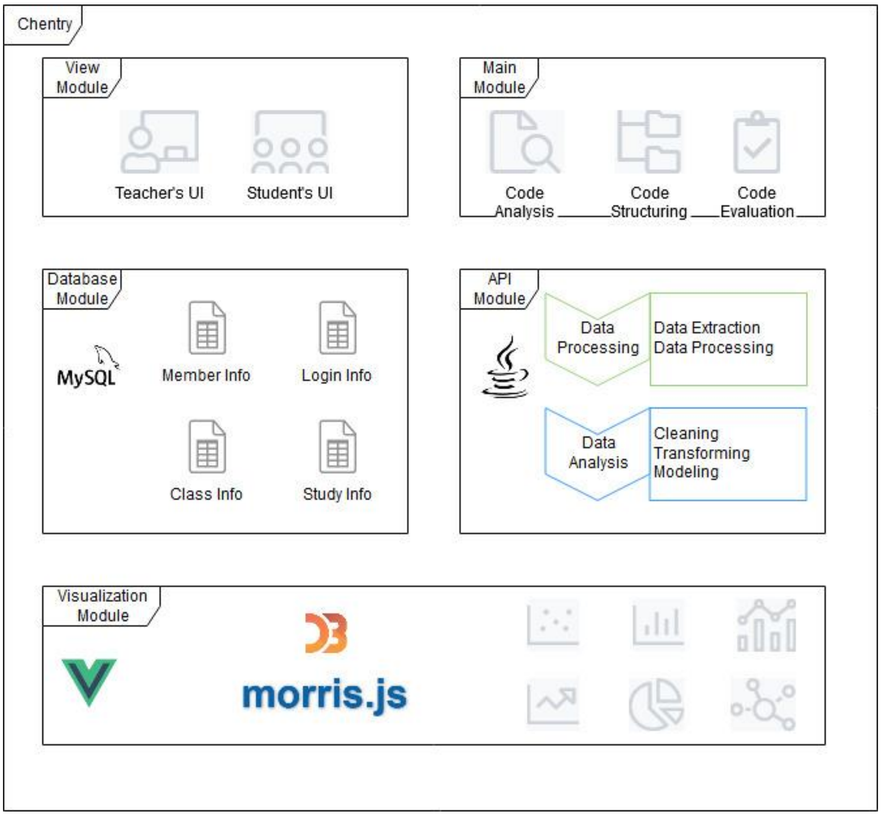


Fig. 1. Overall structure of visualization system

4 Data Visualization

In this section, we explain our data visualization system. In previous sections, we noted what Chentry system collects a large amount of data. But it's not enough to just store the data. Because data must be used to be valuable. In order to gain valuable information from the data, we develop a data analysis system based on Chentry data. Our system can be used by admins, teachers.

4.1 Admin

Admins show visualized information of teachers, classes, and students. Admins can monitor information like areas of teachers, genders of members, new members by months, measurement of all data in current time, etc.

In Fig. 2, the number of member in areas is shown. The number of members is displayed in a map using a GeoJSON file. By seeing this graph, the spread of the members who use our system is understood.

Gender of members is important information to understand the attitude of members. For instance, we can see which gender is more interested in block-based programming topic. The number of members in gender is shown in Fig. 3.

The system shows members (teachers, students) information including their gender, school type, etc. Fig. 4 is the visualization of the number of new members, classes in months.

In Fig. 5, We can see the teachers who have most classes. Based on this graph, we can understand the teaching load.

We work with several elementary and middle schools. In this system, admins can understand information about schools, teachers, and students. For instance, Fig. 6 shows the number of teachers, students from schools.

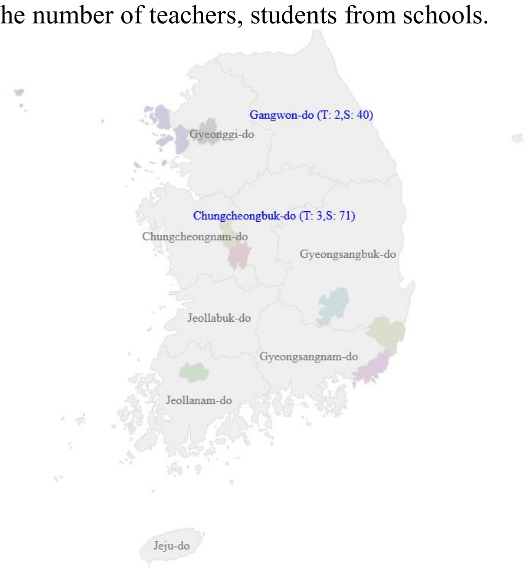


Fig. 2. Number of members by area

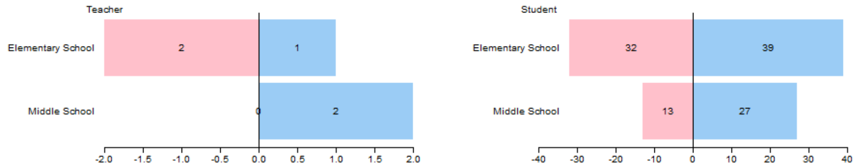


Fig. 3. Number of members by gender

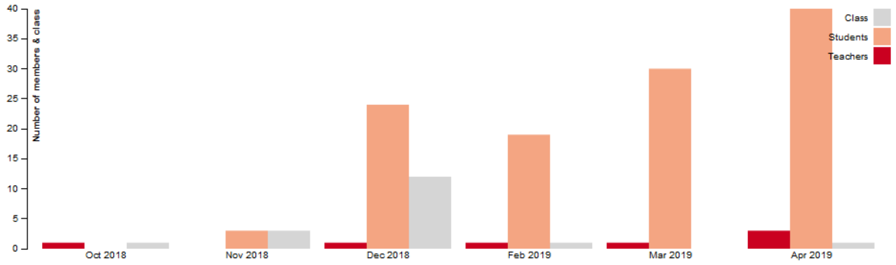


Fig. 4. Number of teachers, students, and classes in months

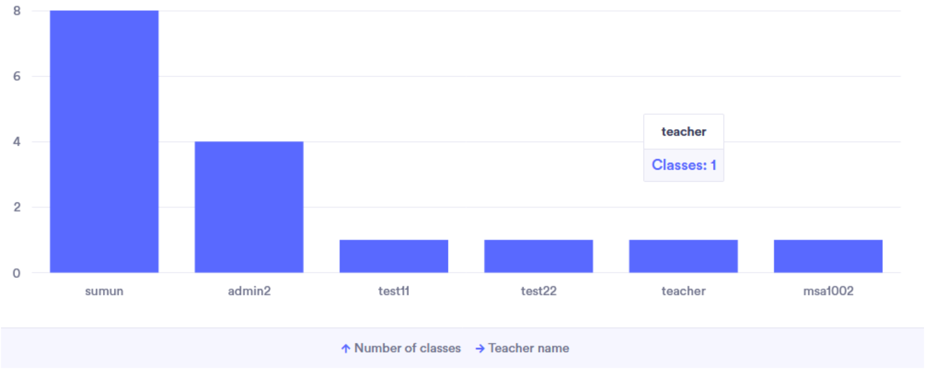


Fig. 5. Number of classes of teachers

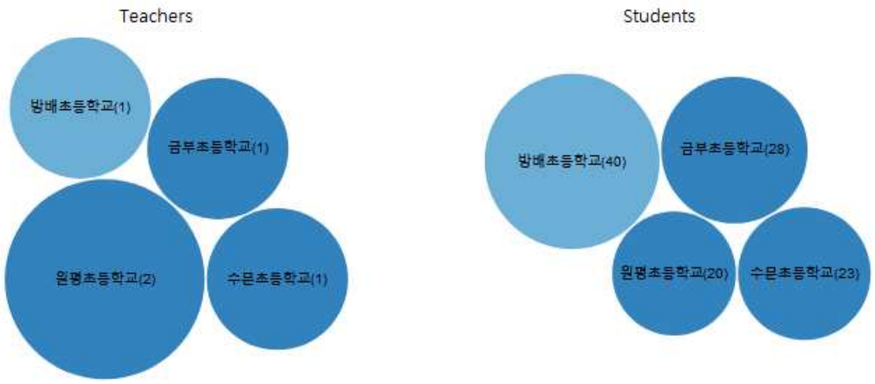


Fig. 6. Number of teachers, students of schools

4.2 Teacher

In this system, Teachers can monitor classes, students, and assignments. The system helps teachers to watch students' progress on assignments and detect the students who need more help from teachers.

Fig. 7 shows student activity in a class. Based on students' behavior, teachers can decide how to deal with his/her students.

In Fig. 8, teachers see progress on a task. By seeing this graph, teachers can know effort on tasks.

In fig. 9 shows a comparison of students progresses to teacher's correct answer. From the graph, Teachers can monitor students' performance of tasks.

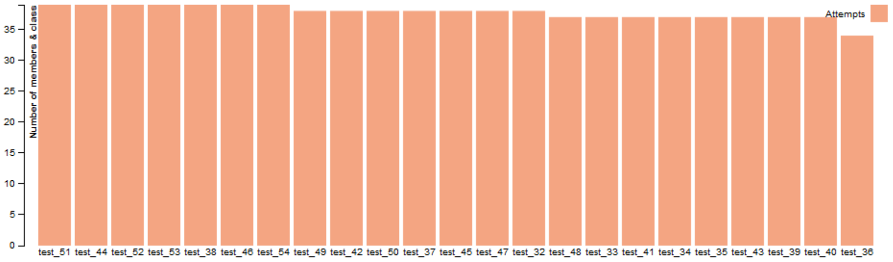


Fig. 7. Number of attempts of student in class

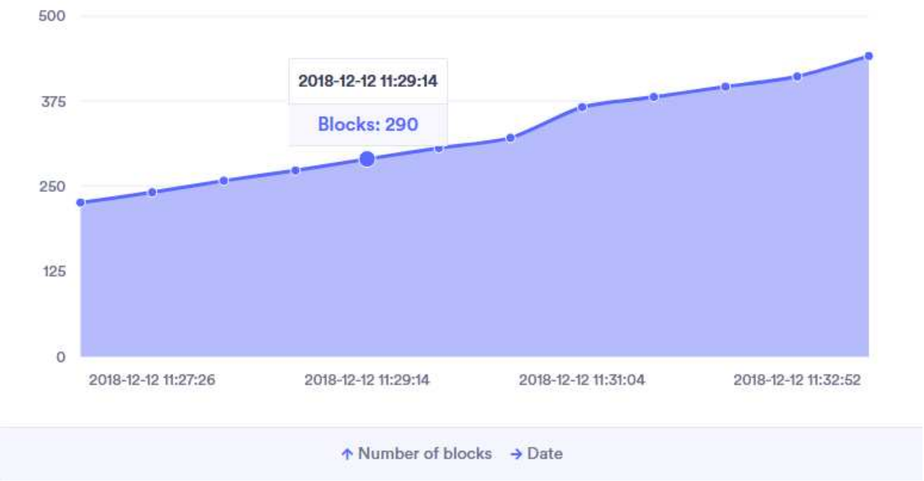


Fig. 8. Progress of students

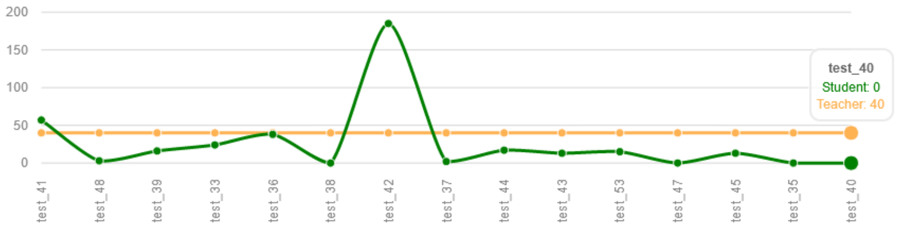


Fig. 9. Comparing students' progress with teachers' answer

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science, ICT & Future Planning) (No.205 2017R1A2B1012909).

5 Conclusion

In this paper, we introduced a data visualization system for assessment of block-based programming. Our system helps admins and teachers to understand the collected data. Based on the visualizations, admins acquire valuable information that can be the right way to develop the system. Also, teachers can conduct his/her students more efficiently. If teachers can understand students' behavior, they can help them in the right way.

In the future, In the future, we will obtain more valuable information using AI. Also, we are going to improve the system using Hadoop.

References

1. Kim JH., Choi JH., Shadikhodjaev U., Nasridinov A., Song KS.: Chentry: Automated Evaluation of Students' Learning Progress for Entry Education Software. In: Lee W., Leung C. (eds) Big Data Applications and Services 2017. BIGDAS 2017. Advances in Intelligent Systems and Computing, vol 770. Springer, Singapore (2019)
2. Jeon, Inseong & Song, Ki-Sang.: The Effect of Learning Analytics System towards Learner's Computational Thinking Capabilities. In: ICCAE 2019 Proceedings of the 2019 11th International Conference on Computer and Automation Engineering, pp 12-16 (2019)
3. Gresse von Wangenheim, Christiane & Hauck, Jean & Faustino DEMETRIO, Matheus & PELLE, Rafael & Alves, Nathalia & BARBOSA, Heliziane & Felipe AZEVEDO, Luiz.: CodeMaster – Automatic Assessment and Grading of App Inventor and Snap! Programs. In: Informatics in Education. vol. 17, pp. 117-150. 10.15388/infedu.2018.08. (2018)
4. Price, Thomas & Dong, Yihuan & Lipovac, Dragan.: iSnap: Towards Intelligent Tutoring in Novice Programming Environments. In: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, pp. 483-488. (2017)
5. Price, Thomas & Barnes, Tiffany.: Comparing Textual and Block Interfaces in a Novice Programming Environment. In: Proceedings of the eleventh annual International Conference on International Computing Education Research. pp. 91-99 (2015)
6. Price, T.W., Dong, Y., & Barnes, T.: Generating Data-driven Hints for Open-ended Programming. In: Educational Data Mining (EDM). pp. 1-8 (2016)
7. Guo, Ming-Li & Wu, Min-Hua & Luo, Li-Ming: A Real-Time and Multi-Dimension Data Types Aware Visualization Technology in Personalized Learning System. In: Proceedings of the 2017 International Conference on Wireless Communications, Networking and Applications, pp. 122 – 126. (2017)
8. Wei Chen, Fangzhou Guo, and Fei-Yue Wang: A Survey of Traffic Data Visualization. In: IEEE Transactions on Intelligent Transportation Systems, vol. 16, pp. 2970-2984 (2015)
9. Gómez-Romero, Juan & Molina-Solana, Miguel & Oehmichen, Axel & Guo, Yike: Visualizing large knowledge graphs: A performance analysis. In: Future Generation Computer Systems, vol. 89, pp. 224-238 (2018)
10. Entry, https://playentry.org
11. D3.js, https://d3js.org/
12. Jersey, https://jersey.github.io/

Customer Interest Detection using Surveillance Videos for Marketing

Jae Jun Lee[1], U Ju Gim[1], Jeong Hun Kim[1], Young-Ho Park[2],
and Aziz Nasridinov[1],

[1] Chungbuk National University, Cheongju, 28644, South Korea,
[2] Sookmyung Women’s University, Seoul, 04310, South Korea
{leejj, kwj1217, etyanue, aziz}@cbnu.ac.kr,
yhpark@sm.ac.kr

Abstract. In many retail vendors, such as Walmart and Costco, recommendations are made through analysis of customer's purchase history. However, it is difficult to identify a customer's interest in a product in real-time. On the other hand, we can use CCTV cameras to model the customer's interest based on his behavior. In this paper, we propose a method for customer interest detection using surveillance vides in the market environment. For this, we first extract customer's keypoints using OpenPose and then, determine if the customer is interested in a product by computing the relationships between keypoints. The proposed method enables us to make customized recommendations and help shops to develop better marketing strategies.

Keywords: Computer vision, surveillance video, customer interest, marketing

1 Introduction

Customer interest is valuable information for marketing. Because it intuitively represents the product the customer wants, it can be an effective marketing method by identifying potential customers. Therefore, various methods to identify customer’s interest have been proposed. For example, large retail vendors like Walmart and Costco analyze customer purchasing patterns through purchase history to identify customer’s interest. At the same time, the products that are likely to be bought can be arranged near to each other. Then it can be set up marketing strategies such as discounts.

Traditionally, the customer's credit card and payment transactions from the POS (point-of-sale) system [1] 's payment transaction data used to collect purchase history for identifying customer's interest. However, these transaction data can not determine how much interest in the product customers have other than what they have purchased. In other words, products that the customer didn't buy but are interested in can never be identified [2]. This interest information can play an important role to understand the customer's interest in the product, and missing it can be a substantial loss for the retailer. On the other hand, we can analyze customer's behaviors through surveillance videos and determine his interests. For example, when the customers have an interest in the product, they tend to stay and look at the product.

A customer with interest in the product tends to stay and look around. Conversely, when there no interest, the customer does not remain at place and passes the shelves. Thus, we identify the customer's interest using the following behaviors: gazing and walking. In this paper, we propose a method for customer interest detection using surveillance videos in the market environment. We collected in-store consumption behavior videos from YouTube. To analyze the customer's behaviors from collected videos, we first extract the customer's keypoints using pose estimation methods, OpenPose [3]. To identify gazing and walking behaviors, we calculated the distance between right and left ankle of a customer. When the distance measure does not change, we can consider it as gazing. On the opposite side, if the distance measure continually changes, we consider it as walking behavior.

The rest of the paper is organized as follows. Section 2 discusses related research. Section 3 explains the proposed method. Section 4 explain experimental results. Section 5 discusses the conclusion and future plans of the paper.

2 Related work

In this section, we explain preliminaries for this research and existing research about customer interest analysis. Section 2.1 explains OpenPose that is mainly used in the proposed method. Section 2.2 discusses the latest research on marketing method through customer interest behavior analysis.

2.1 OpenPose

This subsection explains the OpenPose that is used to understand the customer's interest. Zhe Cao et al. [3] proposed a method that estimates 2d human pose from a video using deep learning in real-time. First, this method learns a model called Part Affinity Fields (PAFs) that individually connect parts of the body in the image. Also, runtime performance and accuracy have been improved by proving that PAF refinement is more critical than combined PAF and body position improvement. Next, a first combined body and foot keypoint detector is presented based on annotated foot datasets internally. It does not only shorten the inference time but also shows that the accuracy is improved individually. Finally, the study has made open-sourced this work as OpenPose, first real-time system to detect body, face, hand, and foot keypoints. This library is now widely used in many research topics such as human behavior analysis and Human-Computer Interaction.

2.2 Existing studies

Liu, J et al. [4] proposed a method that classifies customer's interest behaviors using surveillance video. The authors considered the following actions: no interest, viewing, turning the body to shelf, touching, picking, returning to shelf, and picking and putting into the basket. These behaviors are used to recognize customer's interest behaviors using the head, body orientations, and arm actions. The head and body orientations are divided into eight directions to estimate whether the customer is looking at the shelves or not. Then, Semi-Supervised Learning is used to optimize the training dataset and classify accurately. Further, Combined Had Feature (CHF) that includes moving arm's position and trajectories is extracted. After extraction, CHF is classified into different kinds of arm actions through the Dynamic Bayesian Network (DBN). Finally, performance and effectiveness were demonstrated through an experiment using measuring the accuracy of head and body orientation and arm actions.

This method is most similar to our method, but this method does not consider “gazing” and “walking” behaviors. Therefore, it is not possible to know where and how much times the customer had an interest in the product. In this study, we propose a method that can visualize customer's interest behaviors in real-time and analyze effectively using two behaviors not considered in the above research.

3 Proposed method

This section explains how to analyze customer's interest behaviors to identify customer's interest. The proposed method consists of a pose estimation step using OpenPose and the keypoints calculation method for analyzing customer's interest behaviors. Section 3.1 explains how to extract and use keypoints through OpenPose of pose estimation methods. Section 3.2 describes how to calculate keypoints of customer's interest behaviors.

3.1 Pose estimation

This subsection explains the pose estimation step of the proposed method. As shown in Fig.1, before pose estimation, the surveillance video is input and divided into the corresponding frames according to the video's default settings. Here, for each frame's image, keypoints are extracted through OpenPose, as shown in Fig. 1 (a). There are many pose estimation methods [5, 6]. However, we selected OpenPose due to its high accuracy. In this paper, the keypoints consist of 18 human joints corresponding to the public dataset COCO [7]. For the sake of simplicity, we consider the interest of one customer. Next, x and y coordinates of the extracted keypoints for each frame are stored.

3.2 Modeling of customer’s interest

This subsection explains the calculation methods for analyzing customer's interest behaviors to identify customer's interest in the proposed method. The x and y coordinates, and confidence of the left and right ankles of total 18 keypoints, including the ankles, are used. Here, confidence means the difference between extracted data with OpenPose and the ground truth data. In this step, first, as shown in Fig.1 (b), the point (C_ankle) is calculated as the median of the x and y coordinates of the left and right ankles. Next, as shown in Fig.1 (c), the distance (A_distance) between C_ankle for each frame is calculated. At this time, calculations are excluded for coordinates whose ankle confidence is below 0.8 for x and y coordinates. In other words, if incorrect ankle points are extracted from keypoints extracted through OpenPose due to occlusion and recognition errors, incorrect results may be output. Additionally, if there are no ankle keypoints out of the 18 keypoints obtained through OpenPose, A_distance is stored as 0 and as "no keypoints" error message. Then the A_distance is normalized to a value between 0 and 1. Here as shown in Fig.1 (d), if the value of A_distance is more than 0.1 is classified as walking, and less than 0.1, then it is classified as gazing behavior. Finally, the customer's interest behaviors calculated according to the customer's ankle keypoints are output in real-time, and A_distance is expressed as a line graph. Through this, it can be identified when and where the customers have interested.

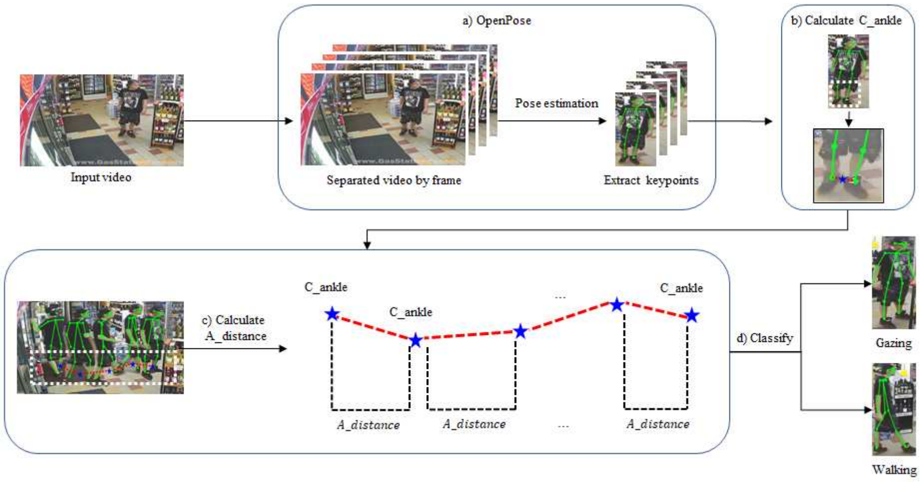


Fig. 1 Overview of proposed method for customer interest detection

4 Experiment

This section explains the experimental settings and results. Section 4.1 describes the experimental settings and data preprocessing method. Section 4.2 explains the results of running the system on in-store consumption video.

4.1 Experimental settings

This section explains an experimental environment, and data preprocessing. The experiments were run on a machine with a single core (Intel Core™ i5-6600 3.30GHz) and 8 GB memory. The proposed method used in the Python programming language.

Before using the 2D keypoints, we performed the data preprocessing. First, if there are several people in the video or if the object is recognized as a person, two or more keypoints group are output. In this experiment, only one person is applied, and others were excluded. Then, if the keypoints for ankles are not extracted correctly, they are excluded from the calculation. Inaccurate keypoints can cause abnormal results.

4.2 Experimental results

This section explains the results of running the system on in-store consumption video using the graph and video output. Fig. 2 shows the system running on in-store consumption video obtained via YouTube. As shown in Fig.2, the three rectangles in the top-left graph represent walking behaviors, and the top-right video visualizes the behavior itself. Considering that the moving distance changes significantly, the amount of change in A_distance can also be seen as significant. On the contrary, as shown in Fig.2, the two rectangles in the left middle graph represent gazing behaviors, and the right middle video visualizes the behavior itself. Considering that the moving distance hardly changes, the amount of change in A_distance is also small. Finally, as shown in Fig.2, the one rectangle in the bottom left graph represents the results of keypoints detection failure with 0 value. It is because the extraction of ankle's keypoints fails with OpenPose, or because the ankle's confidence is low and inaccurate, it is excluded from the calculation.

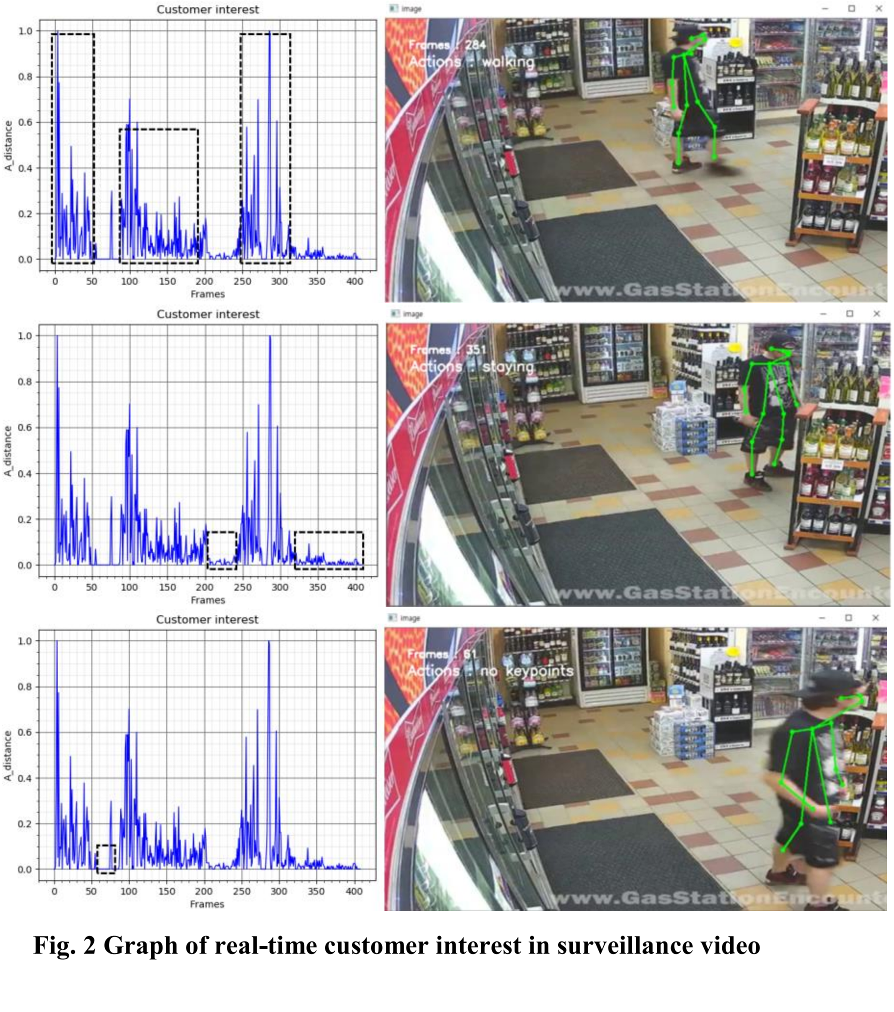


Fig. 2 Graph of real-time customer interest in surveillance video

5 Conclusion

In this paper, we have proposed a method for analyzing customer's interest behavior through surveillance video. Recently, many large retail vendors analyze customer's preferred products and use them for marketing. However, these methods are not suitable to obtain customer's behavior in real-time as it mainly relies on past purchase data of customers. We experiment with each customer's interest behavior by applying the system to the in-store consumption video and explain why the proposed method has the advantage. In the future, we will not only identify customer interests through gazing and walking behaviors analysis but also detect specific products. It will help for enhancing the analysis of customer interest in specific product. In addition to gazing and walking behaviors, touch behaviors can be added to identify customer's interest with higher accuracy.

Acknowledgments. This work was supported by the Institute for Information and Communications Technology Promotion (IITP) Grant funded by the Korean Government (MSIT) (SIAT CCTV Cloud Platform) under Grant 2016-0-00406.

References

1. Reeder, Kenneth Rodney.: Point of sale method and system.In: U.S. Patent No. 6,014,636. (2000)
2. Abdel-Basset, M., Mohamed, M., Smarandache, F., and Chang, V.: Neutrosophic association rule mining algorithm for big data analysis. In: Symmetry 10(4):106 (2018)
3. Zhe Cao., Gines Hidalgo., Tomas Simon., Shih-En Wei., and Yaser Sheikh.: OpenPose: realtime multi-person 3D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018)
4. Liu Jingwen., Yanlei Gu., and Shunsuke Kamijo.: Customer behavior classification using surveillance camera for marketing. In: Multimedia Tools and Applications 76.5 6595-6622 (2017)
5. X. Chen and A. Yuille.: Parsing occluded people by flexible compositions. In: Computer Vision and Pattern Recognition (CVPR), (2015)
6. X. Chen and A. L. Yuille.: Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Advances in Neural Information Processing Systems (NIPS), pages 1736–1744, (2014)
7. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick.: Microsoft coco: Common objects in context. In: ECCV, (2014)

A Study on the Distribution Blockchain DApp Using IP-USN

Seong-kyu Kim*
*CEO/Ph.D of GOB Universal PTE., LTD, Singapore..
*Author Email: guitara7@skku.edu or guitara77@gmail.com

Abstract. This paper has various methodologies for activating the service and the user's use behavior regarding the rapid use of IP-USN, which has emerged as a form of the digital convergence trend that has recently been rapidly progressing in society as a whole. However, the resulting security problem is becoming serious. Analyze blockchain, an IP-USN's concept, service characteristics, and security system that is essential in analyzing this, and analyze the behavior of users for distribution in relation to the use of the application services of the relevant dApp, and derive strategic implications. To this end, when using blockchain technology to mount blockchain technology using IP-USN network, blockchain technology is applied against existing IP-USN to enhance security strength and derive a more secure model by trimming fintech technology to continuously develop and further enhance the global logistics network in the future, so that the technology can be introduced and present to life as an Internet of Things). Chapter 1 described the overall introduction and Chapter 2 described blockchain and IP-USN. Chapter 3 introduces a model that incorporates blockchain technology into the distribution network through the actual IP-USN network, and Chapter 4 presents distribution, logistics, blockchain. Chapter 5 concludes with a conclusion.

Keywords: Blockchain, Bigdata, Artificial Intelligence, Smart Contract, IoT

1. Introduction

In the era of ownership since the Industrial Revolution, where importance to physical space has been increasing, and years after the 20th century IT Revolution, the era of sharing through access to electronic space is coming, a new environmental and miracle paradigm shift based on innovative technological evolution is taking place. This is the emergence of ubiquitous space to maximize interspace convergence and harmony and minimize collisions and breaks in space [1-3]. IP-USN technology, combined with the existing advanced information and communication technology, is expected to be an essential component technology for realizing this ubiquitous reality and will be located throughout social and cultural life. In addition, IP-USN technology is expected to have a significant impact on the IT industry as well as the manufacturing, distribution, and services of national, procurement, construction, transportation, and logistics sectors as well as the non-IT industries in general, and it is expected to exist as an environment encountered in our lives rather than simply as a means of communication networks. Such IP-USN technology is expected to be widely used throughout society due to its convenience, best-of-breed, and efficiency characteristics. Reflecting these characteristics, countries around the world, including the U.S. Department of Defense, the Science Foundation, the Japanese Ministry of General Affairs, and the European Framework Program, have implemented policies to promote national interest in IP-USN. As a result, the IP-USN market predicts steady growth of 25.5% per year on average in terms of global market size, according to IDTechEx/ETRI projections. However, the current IP-USN market appears to be in place at the last test, which is just before commercialization. Global retailers such as Wal-Mart, Target, Tesco, Metro and Albetson have announced the deployment of their consumer supply chain IP-USN systems, and the Pentagon has announced similar IP-USN projects [4-10].

Many major global companies have begun and appear to be considering implementing sit-down assessments and striking out marketability in order to build IP-USN technologies and systems according to their respective circumstances, with the Food and Drug Administration (FDA) currently calling for automatic recognition systems to be deployed in the drug-command. Thus, users can deploy an IP-USN solution under the global IP-USN standard, a viable business model, and a proven ROI, and expectations for the UHF approach are still very high at various stages of system deployment. Given suppliers' obligation to deploy IP-USN systems, the long-term future of UHF-based solutions is expected to be bright in terms of sales and margins.In this distribution, security vulnerabilities still exist, despite the proliferation of technology based on a network of highly skilled technologies [11-14]. Therefore, this paper aims to overcome the security vulnerability of IP-USN networks by using blockchain technology and further demonstrate blockchain technology with good performance and security.

2. Related Research

2.1. Blockchain

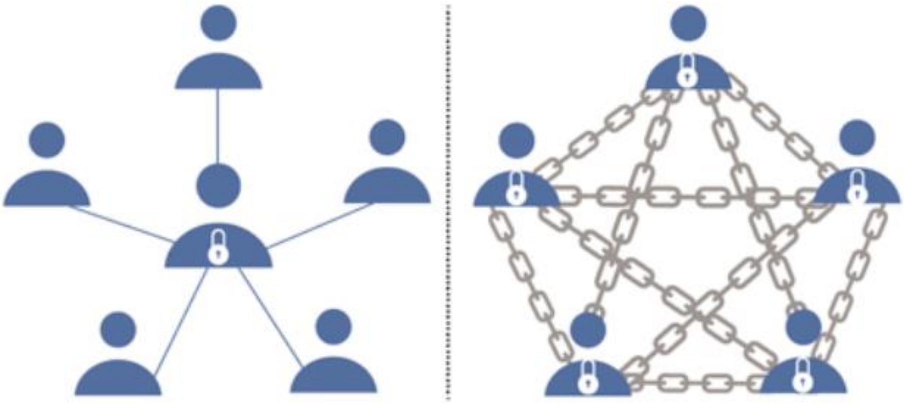
A new database with the ability to keep data spread across geographically separated servers, so that recorded data is not lost (non-changeable) and that some servers continue to operate (allow Byzantine failures)[15-18].

A data storage unit called a block is generated in a constant time period and is characterized by having a data verification model held between each server called a consensus algorithm. Bitcoin is the first blockchain application to operate on a public blockchain accessible to anyone, and with the advent of a highly secure database, it has been recognized as a virtual currency for its high trust in recording the number equivalent to transactions and balances, which continues to withstand various attacks.

The blockchain has the first non-changeability. Unchangeability means that each transaction (data) is stacked on a continuous block. These blocks are dependent on each other, and if some of the past data is changed, all subsequent transactions need to be changed to a coherent form, which is virtually impossible to change.

Allow second Byzantine disorder. This means that blockchain operates normally even if there are a certain number of Byzantine nodes (malicious nodes or faulty computers). Third, remove the single point of failure (SPOF). Elements that cause the entire system to crash if a single point is not functioning. In the system so far, masters, controllers and certification authorities are the single obstacles. There is no single barrier to blockchain [19-22].

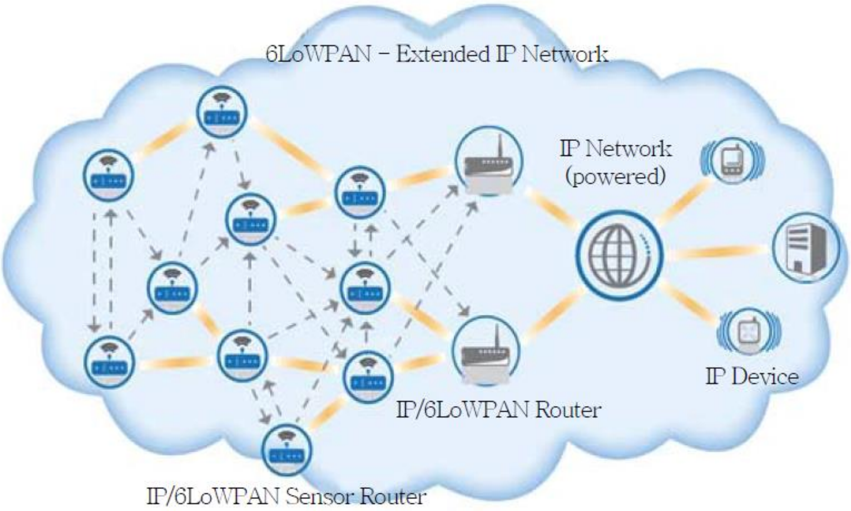
Blockchain is also a distributed computing technology-based data falsification prevention technology based on distributed data storage where small data called blocks are stored in a chain-type distributed data storage environment created based on the P2P method, so no one can be modified arbitrarily and anyone can see the results of the changes. This is essentially a form of distributed data storage technology, designed to prevent arbitrary manipulation by operators of distributed nodes as a list of changes that record continuously changing data on all participating nodes. A well-known example of the application of blockchain is bitcoin, a decentralized electronic book that records the transaction process of cryptocurrency [23-27]. The transaction records operate on computers that are required to encrypt and run blockchain software, and most cryptocurrency, including bitcoin, are based on blockchain technology. shown in [Fig. 1].



[Fig. 1]. Blockchain Architecture

2.2. IP-USN

IP-USN (Internet Protocol-Ubiquitous Sensor Network) refers to a technology that guarantees broad scalability and mobility by integrating USN networks such as sensor nodes, gateways and sink nodes based on existing IP infrastructure. IP-USN is a next-generation core technology that can form sensor networks and provide variety of services in a desired location by connecting with Internet infrastructures such as BcN (Broadband Integrated Network), Next Generation Internet Address System (IPv6), WiBro, and wireless LAN. It is implemented by combining IP (IPv6) into a low-power wireless sensor network (IEEE 802.15.4) network and has the advantage of being directly linked to existing Internet services [28-29]. Competitive technologies include ZigBee, a near-field wireless technology. If a zigbee is suitable for a small sensor network, IP-USN is suitable for large networks such as U-City. IP-USN is expected to be widely used in logistics, disaster prevention, military, home network and U-City. In addition, using IP-USN, the home network service will be able to freely control household appliances through various information devices such as mobile phones and wired or wireless Internet. In addition, telematics can be used to monitor the air pressure of the engine and tire of a driving vehicle in real time to support the safe operation of the vehicle or to provide information on organized traffic, maps, and tourism. Standardization work for IP-USN technology dissemination is also active. The International Organization for Standardization (IETF)'s low-power wireless personal network (6 LoWPAN) working group is pushing for standardization of IP-USN [30-31]. Korea is also actively participating in standardization. The 6LowPAN protocol used in IP-USN is a concept that unlike the Zigbee protocol currently operating on IEEE 802.15.4, sensor nodes are treated as an independent IP node shown in [Fig. 2].



[Fig. 2]. IP-USN Architecture

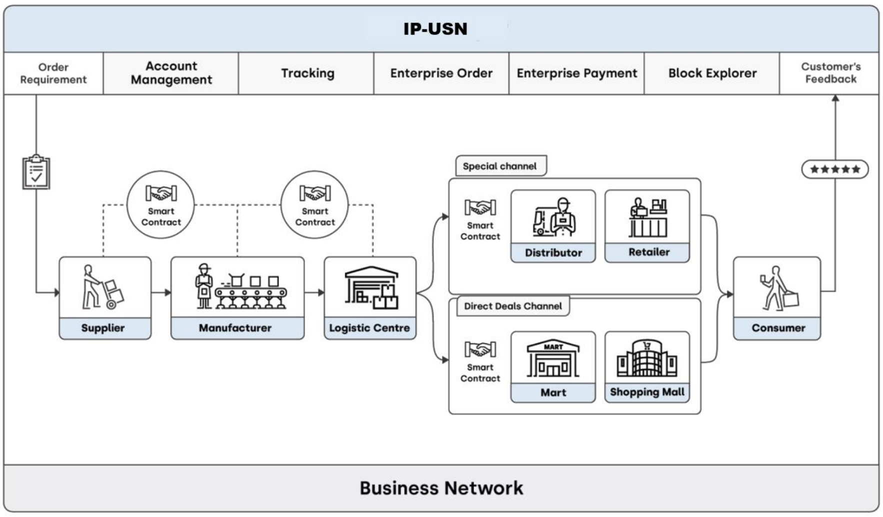
3. Design and Implementation of IP-USN Blockchain

3.1. Issue Raising

IP-USN combines USN's sensor node, gateway and sink node into the IP infrastructure of the existing high-speed Internet, and in short, it combines the Internet into a closed USN that has been sporadically used in disaster prevention, weather and building irrigation. IP-USN can easily form sensor networks in any location by connecting existing sensor networks with internet infrastructures such as BcN (Broadband Integrated Network), next-generation Internet address system (IPv6), WiBro, wireless LAN, etc. and is expected to be widely used in disaster prevention, military, home network, and UCity. The IP-USN coverage is endless. For example, the weather data collection network of the Korea Meteorological Administration or road ice data of the Korea Highway Corporation can be uploaded to the relevant agency's server via the Internet and fed back to the information demander. This IP-USN technology is required for fire departments to detect in case of fire, and for transport cargo, location information is tracked by the national logistics distribution control system. In particular, IP-USN is a USN that is directly connected to the Internet, and its potential is significant in that it will open the way for efficient use of existing Internet infrastructures. It is also possible to develop IP-USN in connection with Internet portals. However, the question is how to organize these networks cost-effectively. And security issues always exist. If security issues arise even if high-speed optical networks are used, it is difficult to use them in current distribution. By winning, the company aims to enhance the distribution system by utilizing blockchain.

3.2. IP-USN Blockchain Architecture

The sensor network has been considered the main application, forming a randomly placed and independent network in any environment to collect the information sensed through communication between the sensor nodes. So we used a lot of low-power MCUs with 8-bit or 16-bit computing power, and we tried a lot to minimize the size and price of the nodes shown in [Fig. 3].

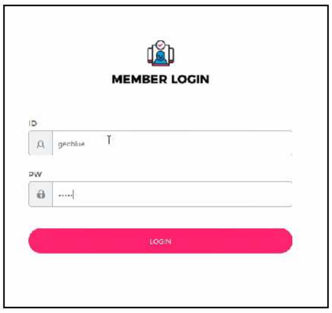


[Fig. 3]. IP-USN Blockchain Architecture Model

However, with no killer applications found, intelligent and large-scale development of application services such as URC robots and LBS is being attempted, instead of services that transmit sensitive data. The requirements are also changing as application complexity increases to consider combining with external network networks, such as IP-USN, and applications placed in defined spaces, such as homes and buildings, such as U-Cities and home networks. And it shows the IP-USN Blockchain distribution architecture.

4. Verifcation of Actual Application Model: Distribution dApp Image Application

A move is underway to apply a block chain, called a distributed public trading unit, to the logistics industry's. In the era of the fourth industrial revolution, technology products such as artificial intelligence, fintech and the Internet of Things are being launched, and blockchain is one of the promising new technologies in the category. The core of the block chain is focused on maximizing information security and record management safety. The blockchain is subtly aligned with the logistics site situation in which the goods services are generated by various stakeholders and processes are operated at various stages. Blockchain acts as a bridge for everyone, including a courier that broadcasts the volume and volume of transactions that were only opened between the shipper and logistics companies, which are parties to the contract, and a subcontractor that is assigned a job. [Fig. 4] is the process of logging into the distribution blockchain.

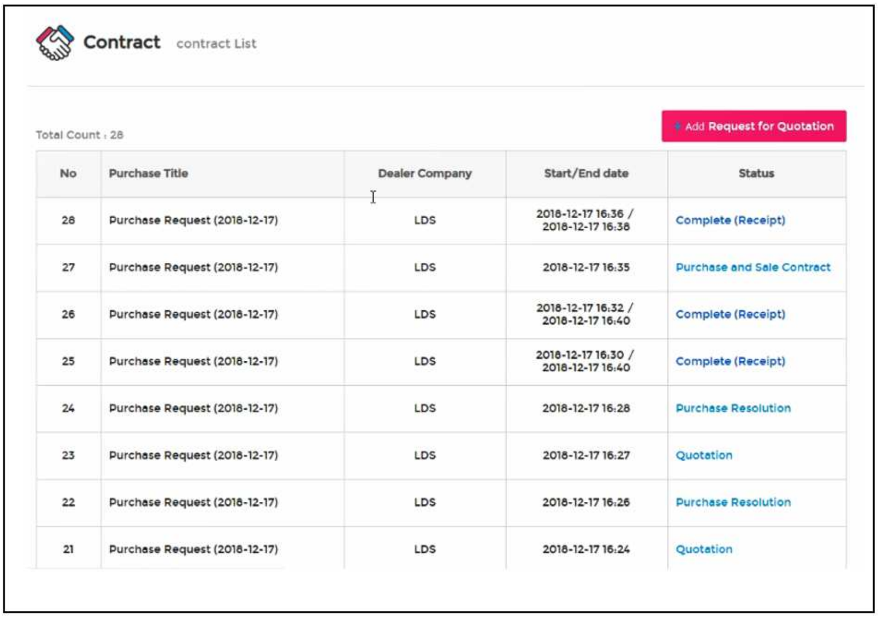


ID: geoblue

PW: ·····

[Fig. 4]. Distribution dApp Model Login

In addition, some are making efforts to incorporate the new technology, called "block chains," into distribution and logistics industry sites. Possible models support the optimization of the SCM as a whole by enhancing the connectivity and reliability of information and processes between stakeholders. In addition, optimization support across the SCM, including areas of shipment, quality control, maintenance/safety environment such as ERP, WMS, and production management systems, is shown through blockchain SmartShift, which strengthens the connectivity of data (information) among stakeholders and enhances the reliability, transparency, and management level of third-party operators and processes [Fig. 5].



Contract contract List

Total Count : 28

No	Purchase Title	Dealer Company	Start/End date	Status
28	Purchase Request (2018-12-17)	LDS	2018-12-17 16:36 / 2018-12-17 16:38	Complete (Receipt)
27	Purchase Request (2018-12-17)	LDS	2018-12-17 16:35	Purchase and Sale Contract
26	Purchase Request (2018-12-17)	LDS	2018-12-17 16:32 / 2018-12-17 16:40	Complete (Receipt)
25	Purchase Request (2018-12-17)	LDS	2018-12-17 16:30 / 2018-12-17 16:40	Complete (Receipt)
24	Purchase Request (2018-12-17)	LDS	2018-12-17 16:28	Purchase Resolution
23	Purchase Request (2018-12-17)	LDS	2018-12-17 16:27	Quotation
22	Purchase Request (2018-12-17)	LDS	2018-12-17 16:26	Purchase Resolution
21	Purchase Request (2018-12-17)	LDS	2018-12-17 16:24	Quotation

[Fig. 5]. Distribution dApp Smartcontract

5. Conclusion

In this paper, IP-USN, a combination of sensors and GPS and other sensing technologies, allows people to share information over a network without feeling that a computer is present, and to collect specific information or provide services.

In other words, sensor networks can be directly linked to society as a whole, from everyday life to industrial sites, by converting the characteristics of objects into electrical characteristics instead of human five senses. Recently, the trend has been linked to U-City's environment, security, disaster prevention and automation It is an area where there is infinite possibility that global markets will grow rapidly and size of global markets will be estimated at about $20 billion in 2018.

In Korea, businesses that have been using near-field technologies such as Zegbee to build sensor networks, use them in various areas such as weather information, bridge management and traffic management have been partially carried out. It is IP-USN that connects these sensor networks directly to IP so that they can be used on the Internet without an intermediate switching device for connecting near-field technology to the Internet. This IP-USN model presents a more secure, more capable model using blockchain technology and a blockchain-based IP-USN model that is not available for public attacks.

References

[1] Nakamoto S.; “Bitcoin: a peer-to-peer electronic cash system,” pp 1-9, 2008.

[2] Jun-Ho Huh,; Kyungryong Seo,; “Blockchain-based mobile fingerprint verification and automatic log-in platform for future computing,” The Journal of Supercomputing, Springer, pp.1-17, 2018.

[3] Seong-Kyu Kim,; Jun-Ho Huh,; “A Study on the Improvement of Smart Grid Security Performance and Blockchain Smart Grid Perspective,” Energies, MDPI, Vol.11, No.7, pp.1-22, 2018.

[4] Yan Chen,; “Blockchain tokens and the potential democratization of entrepreneurship and innovation,” SSRN, pp.12-13, 2017.

[5] Y Nir Kshetri,; “Blockchain’s roles in meeting key supply chain management objectives,” International Journal of Information Management, Elsevier, 80-82., 2018.

[6] Alexander Savelyev,; “Copyright in the Blockchain era: Promises and challenges,” Computer Law & Security Review, Elsevier, 2018.

[7] Jun-Ho Huh,; Sugarbayar Otgonchimeg,; Kyungryong Seo,; “Advanced metering infrastructure design and test bed experiment using intelligent agents: focusing on the PLC network base technology for Smart Grid system,” The Journal of Supercomputing, Springer, Vol.72, No.5, pp 1862-1877, 2016.

[8] Nir Kshetri,; “Blockchain's roles in strengthening cybersecurity and protecting privacy,” Telecommunications Policy, pp 20-23, 2017.

[9] Seong-Kyu Kim, Ung-Mo Kim, Jun-Ho Huh,; “A Study on Improvement of Blockchain Application to Overcome Vulnerability of IoT Multiplatform Security,” Energies, MDPI, Vol.12, No.3, pp.1-29, 2019.

[10] Jun-Ho Huh,; Kyungryong Seo,; “A Typeface Searching Technique Using Evaluation Functions for Shapes and Positions of Alphabets Used in Ancient Books for Image Searching,” International Journal of Hybrid Information Technology, SERSC, Vol.9, No.9, pp. 283-292, 2016.

[11] Richard B. Levin,; Peter Waltz,; Holly LaCount,; “Betting Blockchain Will Change Everything – SEC and CFTC Regulation of Blockchain Technology, Handbook of Blockchain,” Digital Finance, and Inclusion, Elsevier, Vol. 2, 187-212, 2017.

[12] Jun-Ho Huh,; “Server Operation and Virtualization to Save Energy and Cost in Future Sustainable Computing,” Sustainability, MDPI, Vol.10, No.6, pp.1-20, 2018.

[13] Christoph Prybila,; Stefan Schulte,; Christoph Hochreiner,; Ingo Webe,; “Runtime verification for business processes utilizing the Bitcoin Blockchain,” Future Generation Computer Systems, Elsevier, 2017.

[14] Janusz J. Sikorski,; Joy Haughton,; Markus Kraft,; “Blockchain technology in the chemical industry: Machine-to-machine electricity market,” Applied Energy, Elsevier, 234-246, 2017.

[15] Sara Saberi, Mahtab Kouhizadeh, Joseph Sarkis,; “Blockchain technology: A panacea or pariah for resources conservation and recycling,” TTC, pp 15-16, 2018.

[16] Qin, Bo,; et al, “Cecoin: A decentralized PKI mitigating MitM attacks,” Future Generation Computer Systems, Elsevier, 2017.

[17] Huaqun Wang,; Debiao He,; Yimu Ji,; “Designated-verifier proof of assets for bitcoin exchange using elliptic curve cryptography,” TTC, pp 21-24, 2017.

[18] Sabine Löbbe,; André Hackbarth,; “Chapter 15: The Transformation of the German Electricity Sector and the Emergence of New Business Models in Distributed Energy Systems,” Elsevier, pp 287-318, 2017.

[19] Dorri, A.; Kanhere, S. S.; Jurdak, R.; “Towards an optimized Blockchain for IoT,” In Proceedings of the Second International Conference on Internet-of-Things Design and Implementation, ACM, pp. 173-178, 2017.

[20] Pop, C.; Cioara, T.; Antal, M.; Anghel, I.; Salomie, I.; Bertoncini, M. “Blockchain based decentralized management of demand response programs in smart energy grids,” Sensors, MDPI , 18, 162, 2018.

[21] Dorri, A.; Kanhere, S.S.; Jurdak, R.; Gauravaram, P.; "Blockchain for IoT security and privacy: The case study of a smart home," In Proceedings of the 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerComWorkshops), Kona, HI, USA, 13–17 March, 2017.

[22] Underwood, S.; "Blockchain beyond bitcoin," Communications of the ACM, ACM, Vol.59, No.11, pp. 15-17, 2016.

[23] Leiding, B.; Memarmoshrefi, P.; Hogrefe, D.; "Self-managed and Blockchain-based vehicular ad-hoc networks," In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, ACM, pp. 137-140, 2016.

[24] Pass, R.; Shi, E.; Fruitchains,; "A fair Blockchain," In Proceedings of the ACM Symposium on Principles of Distributed Computing, ACM, pp. 315-324, 2017.

[25] Karame, G.; "On the security and scalability of bitcoin's Blockchain," In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, ACM, pp. 1861-1862, 2016.

[26] Kiayias, A.; Koutsoupias, E.; Kyropoulou, M.; Tselekounis, Y.; "Blockchain mining games," In Proceedings of the 2016 ACM Conference on Economics and Computation, ACM, pp. 365-382, 2016.

[27] Hori, M.; Ohashi, M.; "Adaptive Identity Authentication of Blockchain System-the Collaborative Cloud Educational System," In EdMedia+ Innovate Learning, Association for the Advancement of Computing in Education (AACE), pp. 1339-1346, 2018.

[28] Lin, I. C.; Liao, T. C.; "A Survey of Blockchain Security Issues and Challenges," IJ Network Security, ACM, Vol. 19, No. 5, pp. 653-659, 2017.

[29] Kshetri N.; "Blockchain's roles in strengthening cybersecurity and protecting privacy," Telecommun Policy, 41, pp. 1027-1038, 2017.

[30] Jabbar, K.; Bjørn, P.; "Growing the Blockchain information infrastructure," In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, pp. 6487-6498, 2017.

[31] Moshe Babaioff, Shahar Dobzinski, Sigal Oren, Aviv Zohar, On Bitcoin and red balloons. Proc. 13th ACM Conf. Electronic Commerce, 56–73, 2012.

Paprika Purchase Prediction by Using Structured and Unstructured Big Data

HyungChul Rah[1.1], Eunhwa Oh[1.2], Wan-Sup Cho[1.3], Aziz Nasridinov[1.4],
Kwan Hee Yoo[1.4], Yongbeen Cho[2]

[1.1] Department of BigData Convergence,
[1.2] Department of Big Data,
[1.3] Department of Management Information Systems,
[1.4] Dept of Computer Sciences
Chungbuk National University, Cheongju, Korea
[2] Agricultural Bigdata Division, Rural Development Administration, Jeonju-si, Korea
hrah@cbnu.ac.kr, oehoeh0131@gmail.com, wscho@chungbuk.ac.kr, aziz@chungbuk.ac.kr,
khyoo@cbnu.ac.kr, cho0yb@korea.kr,

Abstract. We investigated if the amount of money spent to purchase paprika is correlated with broadcasting news, broadcasting entertaining programs and the social media in which paprika was mentioned between year 2010 and year 2017 in order to facilitate consumption promotion of paprika. In this study, we applied the method to paprika that was used for onion purchase estimation. We identified the statistically significant correlations between paprika purchase with broadcasting entertainment and blogs that mentioned paprika and diet whereas no statistically significant correlations between paprika purchase with broadcasting news. Based on the results, what could be suggested may include a) for promoting consumption of paprika, publicity by using broadcast entertainment programs and blogs appears more effective than news; b) for promoting consumption of paprika through broadcasting programs and blogs, it is necessary to emphasize paprika as diet food. Our results could be applied to promote paprika consumption.

Keywords: Paprika, Agri-Food, Social Media, Purchase Prediction

1 Introduction

It has been reported that various social media and mass media has influenced the agri-food consumption because of the rapid and massive information transmission through social network services (SNS) and broadcast programs especially when food-borne disease and animal disease outbreaks [1]. The impacts of SNS, internet information search, and broadcasting programs on consumer's purchase have been mainly studied in the tourism [2]. Recently, however, its application in the agriculture field has been reported [3].

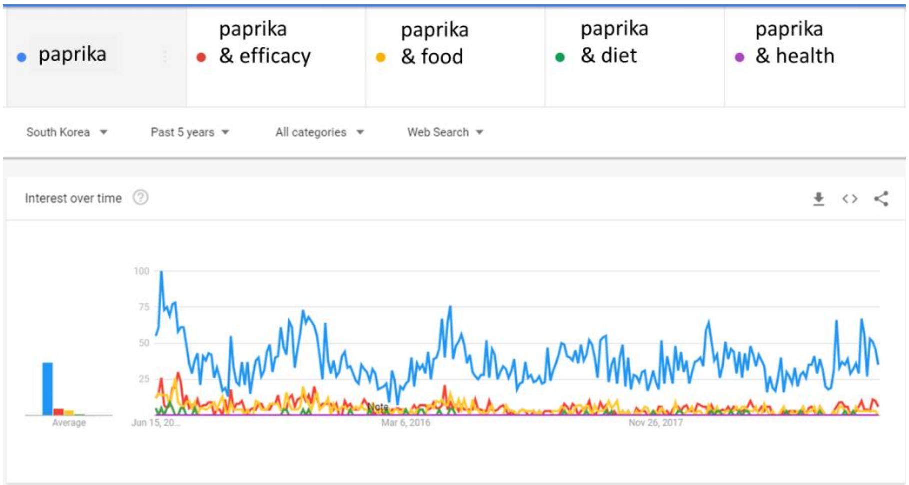
According to the news reports in Korea since year 2010, paprika farming area has increased and economy recession has been prolonged, which led to fall in paprika consumption and paprika price [4, 5]. In order to facilitate paprika consumption, there have been needs to analyze impacts of SNS, internet information search, and broadcasting program on paprika purchase and identify potential factors that can promote paprika consumption in advance.

Based on literature search, no paper has been reported that unstructured data such as social big data was used to forecast paprika consumption. However, we modified the methods used in our previous report on onion purchase prediction in order to identify potential relation between paprika consumption and news, broadcasting programs and SNS [3].

2 Methods

In order to search keywords that are related with paprika consumption, we utilized multiple sources such as Socialmetrics and Google Trend that we used in one of our previous reports [6] . We searched paprika-related words by using Socialmetrics solution of Daumsofts (http://www.socialmetrics.co.kr/) and found that food/cooking related keywords such as ‘sauce’ and ‘pepper’ as well as the name of a one of the Korean idols who ate paprika for her weight loss diet were one of the most frequently mentioned related keywords. We also searched paprika-related words by using Google Trends (https://trends.google.com/) and found that food/cooking related keywords such as ‘how to make paprika salad’ and ‘how to make paprika pickle’ were most frequently mentioned related keywords

Based on the search results of related keywords, we compared query frequencies among four keywords (‘paprika’ ‘paprika & efficacy,’ ‘paprika & food,’ ‘paprika & diet,’ and ‘paprika & health’) over the five years by using Google Trends and found that query frequencies with ‘paprika & efficacy,’ ‘paprika & food,’ and ‘paprika & diet,’ are higher than ‘paprika & health’ (Fig. 2).



● paprika

● paprika & efficacy

● paprika & food

● paprika & diet

● paprika & health

South Korea ▾ Past 5 years ▾ All categories ▾ Web Search ▾

Interest over time

Fig. 1. Search trends of paprika-related keywords for the last 5 years in Google Trend (https://trends.google.com)

Agri-food consumer panel data were collected as structured data in order to estimate the amount of money spent to purchase paprika for the periods of years 2010 and 2017. Unstructured data including Korean broadcasting news, broadcasting entertainment programs and social network (blogs) in which words “paprika,” “paprika and efficacy,” “paprika and food,” and “paprika and diet” were mentioned between year 2010 and year 2017. Collected data were managed in MongoDB as previously described in 2017 [7]. The collected data were transformed into weekly format in order to analyze lagged correlation between paprika purchase (amount of money spent to purchase paprika) and broadcasting data (TV and entertaining programs) and SNS data. Cross-correlation function in R functions was used to estimate lagged correlation.

3 Results

According to the analysis of lagged correlation between paprika purchase and unstructured data, statistically significant correlations with 21 – 23 weeks of lagged time (highest correlation coefficient 0.123 at 23 week) between the broadcasting entertainment programs mentioning words “paprika and diet” with positive terms and the amount of money spent to purchase paprika of the consumer panel data between year 2010 and year 2017 (Fig. 2) were identified.

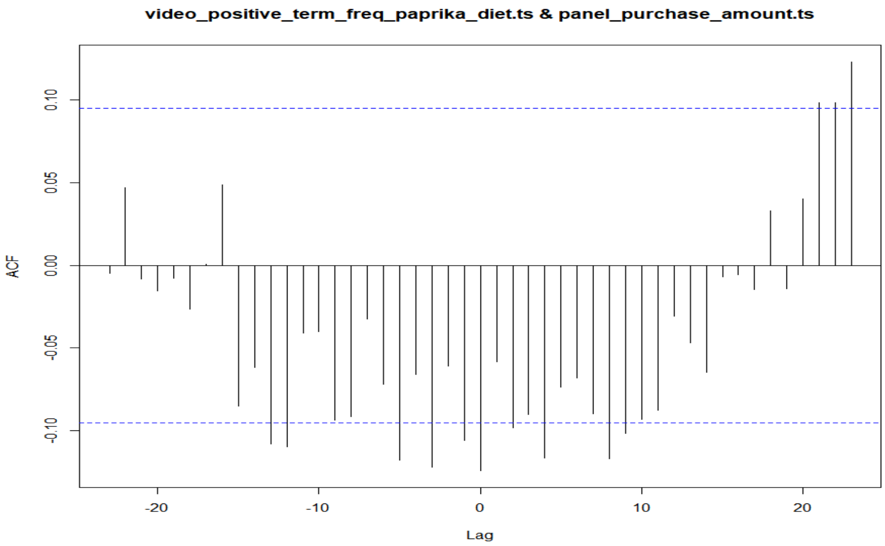


Fig. 2. Cross-correlation coefficients between the broadcasting entertainment programs mentioning paprika and diet and the consumer panel data (paprika purchase) between year 2010 and year 2017

Between blog comments of the blogs mentioning words “paprika and diet,” and the amount of money spent to purchase paprika of the consumer panel data between year 2010 and year 2017, statistically significant correlations with 15 – 18 weeks of lagged time (highest correlation coefficient 0.121 at 17 week) were identified.

However, a statistically significant correlation was not identified between the broadcasting news mentioning words “paprika,” and the amount of money spent to purchase paprika of the consumer panel data between year 2010 and year 2017.

4 Discussion

In our study, we aimed to identify relation between paprika purchase and frequencies of paprika-related keywords in Korean broadcasting news, broadcasting entertainment programs and social network (blogs) in order to facilitate consumption promotion of paprika. We identified the statistically significant correlations between paprika purchase with broadcasting entertainment and blogs that mention paprika and diet with lagged time longer than onion purchase that we previously reported. This long lagged time may come from paprika consumption is for diet, which is optional and it takes time to commit weight loss diet, which are to be confirmed in further studies.

Based on the results, we could suggest the followings:

- When promoting consumption of paprika, publicity by using broadcast entertainment programs and blogs appears more effective than news
- When promoting consumption of paprika through broadcasting programs and blogs, it is necessary to emphasize paprika as diet food and demographics that are interested in weight loss or diet

In this study, we applied the method to paprika that was used for onion purchase estimation [3]. Our results could be applied to promote paprika consumption. However, the trends of paprika purchase that can appear in time-series data could affect the correlations between paprika purchase with broadcasting entertainment and blogs that we identified. It is necessary to carry out the further studies to optimize analysis methods for better estimation of consumer purchase of various agri-foods.

Acknowledgments. This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET), funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (319003-01)

References

1. Shin M-H, Oh S-H, Hwang D-Y, Seo S-S, Kim Y-C. Effect of SNS Characteristics on Consumer Satisfaction and Purchase Intention of Agri-food Contents. JOURNAL OF THE KOREA CONTENTS ASSOCIATION. 2012;12(11):358-67.

2. Park N-H, Kim H-B. The Effect of Keywords of Internet Search Engines on the Demand of Chinese Inbound Tourists: An Application of the Baidu Index Data. Journal of Tourism Sciences. 2016;40(3):159-74.

3. Rah H, Oh E, Yoo D-I, Cho W-S, Nasridinov A, Park S, Cho Y, Yoo K-H. Prediction of Onion Purchase Using Structured and Unstructured Big Data. The Journal of the Korea Contents Association. 2018;18(11):30-37.

4. Ahn H-Y. Plunged Paprika Price. Ikpnews. August 28, 2015.

5. Kim Y-M. Doubled Burden of Paprika Farmers. Agrinet. May 2, 2017

6. Rah H, Park S, Kim M, Cho Y, Yoo K-H. Analysis of Social Network Service Data to Estimate Tourist Interests in Green Tour Activities. International Journal of Contents.2018;14(3):27-31.

7. Rah H, Park K, An B, Choi S, Chae D, Yoo KH. Development of Prediction Model of Agro-Food Demand by Unstructured and Structured Bigdata. The 5th International Conference on Big Data Applications and Services; November 23-25, 2017; South Korea: Korea Big Data Service Society; 2017. p. 122-7.

Prediction of the Production Unit and the Cultivated Area of Onion using the Statistical Methods

Yuha Park[1,1], Myung Hwan Na[1,1], Wanhyun Cho[1], Min Soo Kim[1],
Inchul Jung[2], Yongbeen Cho[3], Deok Hyun Kim[4],

[1] Chonnam National University, 77, Yongbongro, Gwangju, South Korea
[2] National Institute of Horticultural and Herbal Science, 100, Nongsaengmyeong-ro, Wanju-gun, South Korea
[3] Rural Development Administration, 300, Nongsaengmyeong-ro, Jeonju-si, South Korea
[4] Jeollanamdo Agricultural extension & Service, 1508, Senam-ro, Naju-si, South Korea
yuhapark0@gmail.com, {nmh, whcho, mskim}@chonnam.ac.kr,
pfe0524@naver.com, cho0yb@korea.kr, kimdh@jares.go.kr

Abstract. Predicting the production of onion is very important for price stability of crops. To stabilize prices, it is necessary to predict the supply of onion through the production volume of onion. In the supply and demand structure of onion, we can estimate the production volume of onion using the area of cultivation and the number of units of production. In this study, variables that affect the area of cultivation and production units of onions were identified through correlation analysis and estimated using four statistical techniques to compare their predictive power. First, in order to predict the cultivated area of onion, we compare four statistical models using four variables such as the cultivated area of onion and garlic, the wholesale market price and net income of onion farms. According to a comparison of four models, regression analysis using principle component analysis is the best model. Second, in order to predict the number of units of production, we compare four statistical models using four variables such as the length change rates, the highest temperature, the average temperature, the minimum temperature, the relative humidity, the precipitation, and the insolation duration. Based on the comparison of the four models, regression analysis using partial least squares is the best model.

Keywords: Production units and cultivated area of onion, stepwise regression, LASSO, principle component analysis, partial least squares regression.

1 Introduction

Onions are an important ingredient in the diet of Koreans. The Korean government is trying to stabilize prices by selecting pepper, garlic, cabbage, radish, and onion as five

[1] Corresponding author.

sensitive vegetables. Onion has large price variability when compared to other crops. On the other hand, the domestic self-sufficiency rate of onion is about 95%, and most domestic onions are consumed. Since the demand for vegetables is generally assumed to be constant, the cause of price fluctuation mainly comes from supply side [1]. In order to reduce the volatility of onion prices, it is necessary to examine the structure of onion supply and demand. Fig. 1 shows the structure of onion supply and demand.

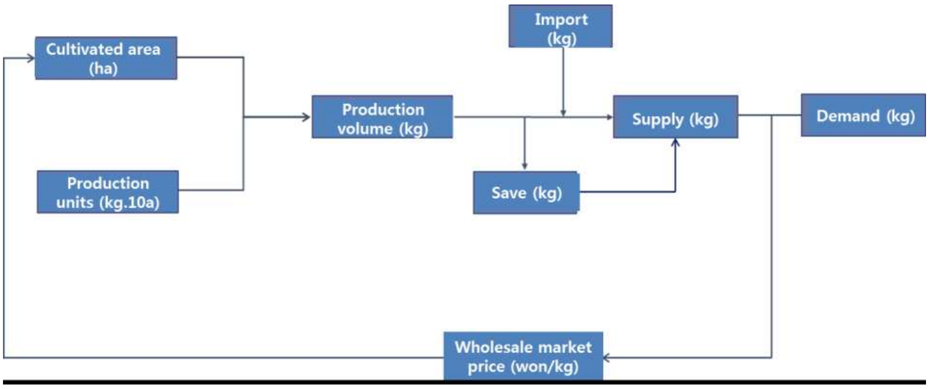


Fig. 1. The structure of onion supply and demand.

Through the structure of onion supply and demand, we can know the production volume (kg) through the cultivated area (ha) and the number of units of production (kg/10a). The number of production units represents unit production per 10 acres of cultivated area. The total supply of onions for a year is obtained by adding the amount of production (kg) to the amount of import (kg). However, in the case of onion, fluctuations in imports account for changes in prices, which are less informative than changes in production [2]. Hence, predicting the production volume of onion is more important to estimate the changes in onion prices. In order to predict the production volume of onion, the cultivated area and the number of units of production are important. When the total supply amount of onion is determined, the wholesale market price of onion is determined under the market economy system and the wholesale market price has a structure that affects the cultivated area of next year onion. The cultivated area can be also determined by autonomous crop selection of farmers. Onion is open field vegetable, and is produced and supplied once a year. The production of onion is greatly influenced by weather change.

In relation to the onion cultivation area, Lee (1996) conducted a correlation analysis of the onion cultivation area of the previous year, the onion price of the previous year, and the factors of the substitute crops for the cultivation area estimation. As a result, although the model was adapted to the model with the cultivation area of the previous year and the annual onion price of the previous year as the explanatory variable, there is a limit to which prediction is not made [1]. Nam and Choi (2015) compared the predictive power by applying the model parameters to the 2-time lag as the explanatory variables for the previous year's cultivation area, wholesale market price, income per 10a, and gross income per 10a. However, this is a limitation on data loss because explanatory variables are applied to different models.

Nam and Choi (2015) used multiple regression models using variable selection method by selecting only significant variables through correlation analysis between monthly production factors and monthly meteorological factors. However, even in Jeonnam, South Korea, the weather information of each region is different, but there has a limitation using the average of the whole region [2]. Lim (2016) used the lowest temperature as a key factor through correlation analysis using weather factors of each mainland as explanatory variables. However, the results of regression models using meteorological factors were not as good as other crops [3]. Choi (2016) was predicted to produce a singular space through the panel models that reflect the spatial information for the panel data. As a result, the sunshine hours (January), the average relative humidity (April), the mean minimum temperature (June), and cumulative precipitation (November) were significant variables. As a result, the change in the actual production number and the estimated value could have a similar result [4].

In order to estimate the total onion production volume, the following procedure is used. First, the estimation of cultivation area of onion: the area of cultivation in the previous year, the price of onion wholesale market and the factors of substitute crops were used as explanatory variables. Variables affecting onion cultivation area are selected by correlation analysis between explanatory variables and onion cultivation area. Using the selected variables, the predictive power was compared after applying it to the four regression models. Second, the estimation of the number of production units of onion: Information collected through direct survey as well as weather information is used as explanatory variables. On the basis of correlation between the number of onion production units and the explanatory variables, we select the variables. Using the selected variables, we compared the predictive power using the same models as the cultivation area estimation.

2 Materials and Methods

2.1 Datasets

In this study, we used datasets for predicting the cultivated area of onion such as the dataset of cultivated area of onion and garlic from 1997 to 2018, the dataset of onion wholesale price from May 1997 to April 2018, and the dataset of annual net income of onion farms from 1997 to 2017. And we used datasets for predicting the number of production units of onion such as the number of production units and cultivated area of onion, onion growth information and monthly weather information (daily average temperature, maximum temperature, minimum temperature, relative humidity, precipitation, solar radiation time). Onion growth information was collected at the actual farmhouse through growth investigation. Growth surveys began on March 1, 2018 and were conducted six times at intervals of 15 days. A total of 55 observations were used as response variables of onion production area. Table 1 shows the datasets used in the study.

Table 1. The datasets consists of largely two parts.

Purpose	Dataset	Period	Source
Cultivated area	Cultivated area of onion and garlic	1997 ~ 2018 (yearly)	KOSIS (http://kosis.kr)
	The wholesale market price	May, 1997 ~ April, 2018 (monthly)	KAMIS (https://www.kamis.or.kr)
	The net income of onion farm	1997 ~ 2017 (yearly)	KOSIS (http://kosis.kr)
The number of production units	The number of production units in the onion production sites	Yearly	KOSIS (http://kosis.kr)
	The growth information	March, 2018 ~ June, 2018 (at the intervals of 15 days)	Collected
	Weather information	Monthly	KMA (https://data.kma.go.kr)

2.2 Methods

In this study, four regression models (stepwise variable selection method, least absolute shrinkage and selection operator, principal component analysis regression, partial least squares regression) are utilized to solve the problems with more variables than the number of observations. When a multiple linear regression model has a linear relationship between response and explanatory variables, and the number of observations is much larger than the number of variables, the least squares estimates can achieve good performance with low variance. However, if the number of observations is not much larger than the number of variables, it can lead to large variations, overestimated model. Also, forecasts for future observations may be poor.

First, the stepwise method is one of the variable selection methods used for multiple regression analysis. This method increases the number of explanatory variables that have the greatest effect on the response variable and increases the number of explanatory variables one by one.

Second, the least absolute shrinkage and selection operator (LASSO) is one of the shrinkage methods. This method is applied to a model that includes all explanatory variables, and constrains the regression coefficients to reduce the size of the estimates of the coefficients. The constraint is to minimize the residual sum of squares in the condition that the sum of the absolute values of the regression coefficients is smaller than the given parametric term.

Third, principal component analysis (PCA) is a method to reduce the dimensionality by finding the principal component of linear combination between variables using the variance-covariance. The principal components represented by the linear combination of variables are independent of each other, and the total variation contained in the p variables is replaced by m ($m \leq p$) principal components. The PCA regression uses the principal component obtained from the principal component analysis instead of explanatory variables. The number of principal components is determined by the number of principal components determined by PCA.

Lastly, the partial least squares method (PLS) is a method of dimension reduction like the PCA. Differences are found in a map-like way when constructing a principal component using response variables. Weights are given to variables that are strongly related to response variables when making the principal component. It is essential that the PLS regression relate the explanatory variables to the response variables through latent factors.

In the study on the cultivation area, we used data from 1998 to 2016 as the train data and compared with 2017 and 2018 as the test data. The root mean squared error (RMSE) values were compared with the predicted values of each model. In the study related to the production units, 50 observations were used as the train data among the 55 observations, and 5 were used as the test data. We compare the performance of each model in the same way as the study on the cultivation area.

3 Experimental results

3.1 Estimation of Cultivated Area

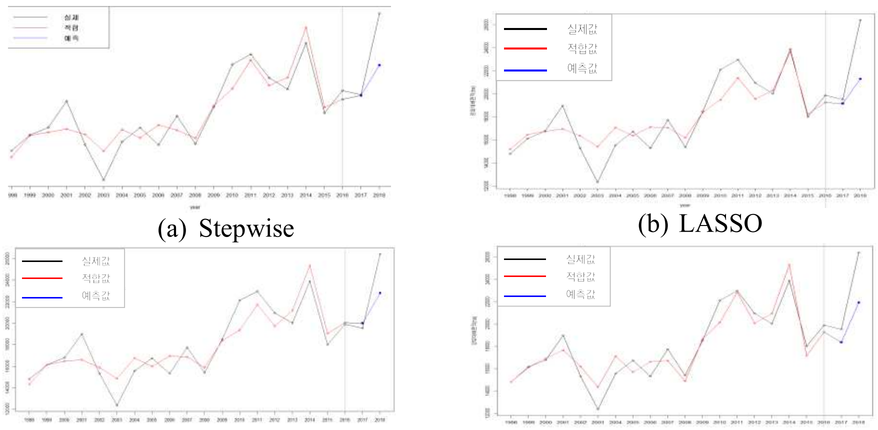
The results of the correlation analysis between onion cultivation area (at time *t*-year) and all explanatory variables showed that the total of four variables, onion cultivation area (at time *t-1*-year), garlic cultivation area (at time *t-1*-year), wholesale market price (From May, *t-1*-year to August, *t-1*-year) and net income of onion farmers (at time *t-1*-year) were selected as statistically significant variables. The results of the four regression analysis are as follows in Table 2.

Table 2. The results of the four regression analysis.

Method	Selected variables	Effect on response variable	R^2 (RMSE)
Stepwise	Onion cultivation area (at time *t-1*-year)	(+)	0.81 (3227.68)
	Wholesale market price (May, *t-1*-year)	(+)	
	Net income of onion farmers (at time *t-1*-year)	(+)	
LASSO	Onion cultivation area (at time *t-1*-year)	(+)	0.83 (3626.28)
	Wholesale market price (May, *t-1*-year)	(+)	
	Wholesale market price (August, *t-1*-year)	(+)	
	Net income of onion farmers (at time *t-1*-year)	(+)	
PCA	PC1 (wholesale market price from June to August, *t-1*-year and net income of onion farmers)	(+)	0.745 (1585.94)
	PC2	(+)	

	(cultivation area of onion and garlic, *t-1*-year)		
	PC3 (wholesale market price on May, *t-1*-year)	(+)	
PLS	PL1 (wholesale market price on July, *t-1*-year)	(+)	
	PL2 (cultivation area of onion, *t-1*-year and wholesale market price on May, *t-1*-year)	(+)	0.75 (3284.47)
	PL3 (cultivation area of garlic, *t-1*-year and wholesale market price on June and August, *t-1*-year)	(-)	

The actual and predicted values using the model are shown in Fig. 2. The coefficient of determination was the highest in the LASSO regression model, but the PCA regression model had the lowest RMSE. Therefore, it was shown that PCA regression model is the best model to predict the onion cultivation area.



(a) Stepwise (b) LASSO

(c) PCA (d) PLS

Fig. 2. Plots of actual values and predicted values of cultivated area of onion in each regression model.

3.2 Estimation of the Number of Production Units

In this study, we used the difference amount of length, length, and temperature related variables such as average temperature, minimum temperature, and maximum temperature in each month as explanatory variables. We also used the relative humidity in each month, monthly rainfall amount in each month, and solar radiation period in each month. The results of the four regression analysis are as follows in Table 3.

Table 3. The results of the four regression analysis.

Method	Selected variables	Effect on response variable	R^2 (RMSE)
Stepwise	Difference amount of length	(+)	0.72 (453.37)
	Monthly rainfall amount in October	(-)	
	Monthly rainfall amount in November	(-)	
	Minimum temperature in November	(-)	
	Solar radiation period in April	(+)	
LASSO	Difference amount of length	(+)	0.75 (508.69)
	Maximum temperature in October	(+)	
	Minimum temperature in November	(-)	
	Minimum temperature in April	(-)	
	Relative humidity in November	(-)	
	Monthly rainfall amount in October	(-)	
	Monthly rainfall amount in November	(-)	
	Solar radiation period in November	(+)	
	Solar radiation period in April	(+)	
PCA	PC1 (Minimum temperature and Relative humidity)	(+)	0.73 (505.34)
	PC2 (Maximum temperature and Monthly rainfall amount)	(-)	
	PC3 (Difference amount of length)	(-)	
	PC4 (Solar radiation period)	(+)	
PLS	PL1 (Minimum temperature and Solar radiation period)	(+)	0.68 (392.81)
	PL2 (Difference amount of length)	(+)	
	PL3 (Length)	(+)	
	PL4 (Maximum temperature and Monthly rainfall amount)	(+)	

The coefficient of determination was the highest in the LASSO regression model, but the PLS regression model had the lowest RMSE. Therefore, it was shown that PLS regression model is the best model to predict the number of production units of onion.

4 Conclusions and Future works

In this study, we estimated the yield of onion by estimating the cultivation area and the number of production units in the supply and demand structure of onion using the four regression models such as stepwise, LASSO, PCA, PLS. First, the regression model using PCA was the best predictive model for forecasting the cultivation area of onion among the four models. Second, regression model using PLS was the best predictive model for forecasting the number of production units among the four models. Prediction of the production volume was possible in mid-April before harvest. As a result, using the proposed PLS model, we can estimate the total production in April of each year before the onion comes out.

In this study, the total cultivation area in South Korea was used as a response variable in predicting the cultivation area. However, the number of production units utilized only using the production number of the main production area. The range of investigation of response variables will be different. In order to predict the production volume, information on the cultivation area of each region seems necessary. In future, we will add information on cultivation area of each region, estimate the cultivation area and the number of production units, and predict the total amount of onion production in each region.

Acknowledgments. This work was supported by the Research Program of Rural Development Administration (Project No. PJ0138672019).

References

1. Lee, J.W.: A Study of Decision-Making Factors of Production for Red Pepper, Garlic and Onions. Journal of Rural Development, 19(3), pp. 27--50 (1996)
2. Nam, K. H., Choe, Y. C.: A study on Onion Wholesale Price Forecasting Model, Journal of Agricultural Extension & Community Development, 22(4), pp. 423--434 (2015)
3. Lim, C.-H., Kim, G. S., Lee, E. J., Heo, S., Kim, T., Kim, Y. S., Lee, W.-K.: Development on Crop Yield Forecasting Model for Major Vegetable Crops using Meteorological Information of Main Production Area. Journal of Climate Change Research, 7(2), pp. 193-203 (2016)
4. Choi, S. C., Baek, J.: Crop yields estimation using spatial panel regression model, The Korean Journal of Applied Statistics, 29(5), pp. 873--885 (2016)

Application of to prediction agricultural/livestock consumption prices based on LSTM

Ga-Ae Ryu[1], seunghyeon Kang[1], Youngbeen Cho[2], HyungChul Rah[3],
Aziz Nasridinov[1], Kwan-Hee Yoo[1*]

[1]Dept. of Computer Science, Chungbuk National University, South Korea
[2]Dept. of Agriculture Bigdata Job in the Planning and Coordination Bureau, Rural Development Administration, South Korea
[3]Dept. of Big Data Collaboration Course, Chungbuk National University, South Korea
garyu@chungbuk.ac.kr, cat1919@naver.com, cho0yb@korea.kr,
{hrah, aziz, khyoo}@chungbuk.ac.kr
[*]Corresponding Author

Abstract. The agriculture and livestock consumption area is not only targeted at consumers of various agricultural and livestock products with unique characteristics but also influenced by various media including unpredictable consumption trends, social atmosphere, events, and accidents. Thus, observations of consumption for agriculture and livestock products from various agencies often show results that are difficult to perform observation and cannot guarantee accuracy. It is easily checked that the observation center often makes the previous observation contents inaccurate. And the center do not support future prices in real time. In this paper, we propose a method to predict the purchase price of agricultural and livestock consumer panels using deep learning by utilizing the collected structured and unstructured data.

Keywords: Agricultural and Livestock Products, prediction, panel purchase amount, Deep Learning, LSTM

1 Introduction

Recently, the prices of various agricultural and livestock products just as pig prices have soared because of African swine fever have been skyrocketing and plummeting, that it is difficult to predict the purchase price of agricultural and livestock products. As a result, many researchers are interested in how to predict the consumption, wholesale, and purchase prices of agricultural and livestock products. The reason for this is that 1) providing information on future prices of agricultural and livestock products makes it possible for market consumers to make rational decisions, and 2) it is possible to the allocation of resources efficiently in a socially. Therefore, if it is possible to predict the consumption price, wholesale price, and purchase price, it can also better suggest the direction of purchase and consumption of agricultural and livestock products from the standpoint of individual and government.

However, the method of predicting the price of agricultural and livestock products is very difficult for various reasons. The reason is as follows. 1) The consumption area of agriculture/livestock products are targeted to various agriculture/livestock products consumers with unique characteristics, 2) it is influenced by various media such as consumer trends, social atmosphere, events and accidents, 3)It is difficult to observe consumption of agricultural and livestock products and the accuracy of prediction is not guaranteed with small amounts of data. (ex. Many agencies, such as the Korea Rural Economic Institute's Agricultural Observation Center, measure agricultural consumption patterns and wholesale prices. If you go to the page and check the contents, you can see the last three months of the transaction price and you can see also only the two or three small amounts historical information of a week earlier than the current date in case of the wholesale price and carry-in quantity. [1])

In this paper, we propose the method of predicting the purchasing price of consumer panel for agricultural and livestock product through deep learning using collected structured and unstructured data (news, blog, broadcast data). [2] The reason for estimating the purchase price of a consumer panel is that it is possible to understand the purchase pattern and consumption price of a consumer depending on the price of the panel. Accordingly, the Long Short-Term Memory [3] method is applied using structured data (such as pig imports, pig purchases, etc.) and unstructured data (related 'pigs' data in news, blogs, broadcasting) from 2010 to 2017 by focusing on 'pigs' in agriculture/livestock products.

2 Related Works

The LSTM (Long Short-term Memory) [3] is a modification of the existing RNN (Recurrent Neural Network) model [4], which is a typical improvement model that mitigates the vanishing gradient problem of RNN [4]. This model includes a memory cell in the hidden node which stores and outputs the value and adjusts the forgetting value. [3] What is interesting in this model is that the LSTM consists of an input gate for the input value, an output gate for the output value, and a forgetting gate for the forgetting value. [3] The learning algorithm of the LSTM uses Backpropagation in the same way as the RNN's [3], the input data is the sequence data, the output data is the output data of LSTM [3]. The LSTM model includes three gates so that the number of weights and the number of biases is about four times that of a typical RNN learning, which means that execution time and learning the time of LSTM model are longer than RNN model's [3]. Despite this, the vanishing gradient problem can be mitigated to obtain more accurate results. [3]

There is a large volume of published studies using LSTM. The study by Tran and Song predicted the water level using deep learning. [5] In that paper, to prevent flooding damage in the city, the authors learned the sequence data of river level observation and predicted water level. RNN [4], RNN-BPTT (Recurrent Neural Network- Backpropagation Through Time) [4], LSTM [3] were used as learning methods. The finding highlights the high performance of the three models, but the LSTM is the most efficient method to find optimal results. [5] In another study, Joo and Choi predicted the stock price fluctuation pattern. [6] The bidirectional recurrent neural network was applied to the learning method through the forward neural network and the reverse neural network. [6] To improve the accuracy of the prediction, the bidirectional LSTM [7] was used in deriving a few errors. As a result, the bidirectional LSTM achieves a higher degree of accuracy than the unidirectional LSTM.

3 Proposed Method

In this paper, we propose the method of predicting the purchase price of the consumer panel using deep learning. The propose method learned and predicted by the LSTM method using sequence data (from 2010 to 2017), which is listed daily. Used data in training are using structured data related to pigs and unstructured data such as news, blogs, broadcasting program that extracted the positive word and negative word count related to pigs.

3.1 Used Data

The data used to predict the purchase price of consumer panels is structured and unstructured data as sequence data from 2010 to 2017. For structured data, it consists of sequence data related to pigs such as retail prices, wholesale prices, purchases, and imports quantity. For unstructured data, we find the contents with a list of pig-related items, and then extract the positive/negative word count from the contents of blogs (Daum, Naver), news (Internet News, Broadcast News) and broadcasting programs (programs related to food). Finally, we calculated the sum of the positive/negative word count result by each day. As a result, we made a total of 2,922 records by combining structured and unstructured data. The data used are shown in Table1.

Table 1. Used Data for predict panel purchase amount

Type	Column	Description
Structured data	panel_purchase_amount_ave	Consumer Panel Purchase Price Average of pigs
	panel_purchase_amount_sum	Consumer Panel Purchase Price sum of pigs
	retail_price_meat	The retail price of pigs
	wholesale_price_carcass	The wholesale price of pigs
	pig_bred_number_quarter_before	Pig breeding in the previous quarter
	pig_slaughtered_number_quarter_before	Pig slaughtered in the previous quarter
	wholesale_price_carcass_quarter_before	Wholesale prices for the previous year quarter
	output_ton_year_before_carcass	Pork production in previous year
	import_ton_year_before	Pork imports in the previous year
	monthly_sales_trend_ton_meat	Monthly pig sales quantity
Unstructured data	video_freq	Video pig refer frequency
	video_positive_term_freq	Video pig positive word frequency

	video_negative_term_freq	Video pig negative word frequency
	news_freq	News pig refer frequency
	news_comment_freq	News comments pig refer frequency
	news_positive_term_freq	News pig positive word frequency
	news_negative_term_freq	News pig negative word frequency
	blog_freq	Blog pig refer frequency
	blog_comments	Blog comment pig refer frequency
	blog_likes	Blog like's frequency include pig contents
	blog_positive_term_freq	Blog pig positive word frequency
	blog_negative_term_freq	Blog pig negative word frequency

3.2 Training and Prediction Method

To predict the purchase price of the consumer panel we apply the LSTM method. Used the data is from 2010 to 2017 and the total number of data is 2922. In this study, we learn data (2,557 records) from 2010 to 2016 and predict and compare data (365 records) from 2017. To predict the data, proceed as follows. 1) data preprocessing 2) correlation analysis and giving weights 3) data training. First, data preprocessing is cleaned the data from 2010 to 2016 and performs data normalization through min-max scaling. Second, the correlation analysis and giving weights has analyzed the correlation of each column with the purchase price of the consumer panel, and give weights according to correlation coefficient results. Finally, data training is used LSTM method to learn data. At this time, we can adjust the factors such as seq_length, rnn_cell_hidden_dim, forget_bias, dropout_rate, learning_rate, training_num, etc., and we can find the best factor after many attempts. The factors are as follows. (seq_length:14, rnn_cell_hidden_dim:20, forget_bias: 0.5, num_stacked_layer:4, dropout_rate:0.5, learning_rate:0.01, training_num:100) After that, we compare and analyze with the 2017 data through the predicted results. Learning and predicting methods are shown in Fig. 1.

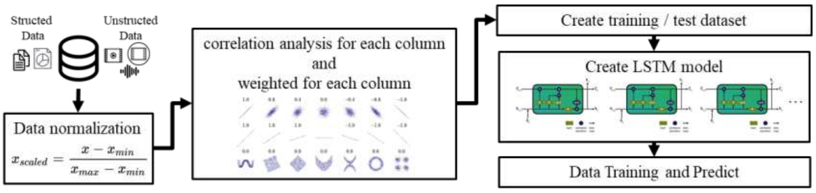

Structed Data
Unstructed Data
Data normalization
$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$
correlation analysis for each column and weighted for each column
Create training / test dataset
Create LSTM model
Data Training and Predict

Fig. 1. Data Training and Predicting method

4 Experimental Result and Future work

In this paper, we propose the method to predict the purchase price of consumer panel for agricultural and livestock products through LSTM by structured and unstructured data from 2010 to 2017. Data learning was used the GPU-based tensorflow [8], and as a result, we could be predicted in 2017 data with 82% accuracy over a 30-second time period. The prediction result is shown in Fig.2.

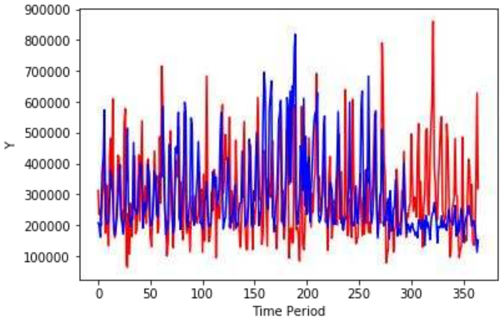


Fig. 2. Predicted Result

Afterward, we will improve the accuracy by applying the adaptive LSTM method for each agricultural and livestock products, and further, provide better service to the users by visualization on the web.

Acknowledgments. This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET), funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (319003-01).

References

1. Korea Rural Economic Institute Agricultural observation site, http://aglook.krei.re.kr/jsp/pc/front/observe/monthlyReport.jsp?ovr_item_code=OVR0000000008&prt_ovr_item_code=OVR0000000008
2. Song, J.H., Kim, J.H., Ryu, G.A., Nasridinov, A., Yoo, K.H.: A Big Data Platform for Collecting Agricultural Data. In: The 4th international conference on next Generation Computing, pp.126--127. Vietnam (2018)
3. Greff, K., Srivastava, R.K., Koutnik. J., Steunebrink, B.R.: LSTM: A search Space Odyssey. In: IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222--2232. (2017)
4. Lipton, Z.C., Berkowitz, J., Elkan, C.: A Critical Review of Recurrent Neural Networks for Sequence Learning. In: arXiv:1506.00019v4, pp. 1--38. (2015)
5. Tran, Q.K., Song, S.K.: Water Level Forecasting based on Deep Learning : A Use Case of Trinity River-Texas-The United States. In: Journal of KIISE, vol. 44, no. 6, pp. 607--612 (2017)
6. Joo, I.T., Choi, S.H.: Stock Prediction Model based on Bidirectional LSTM Recurrent Neural Network. In: Journal of Korea Institute of Information, Electronics, and Communication Technology, vol. 11, no. 2, pp. 204--208 (2018)
7. Schuster, M., Paliwal, K.K.: Bidirectional recurrent Neural Networks. In: IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673--2681 (1997)
8. Tensorflow, https://www.tensorflow.org/

Detection and Identification of Various Insect Pests of Tomato Using Several Deep Learning Algorithms

Wanhyun Cho[1], Myunghwan Na[1], Sangkyoon Kim[2], Hyerim Lee[3,]

[1]Department of Statistics, Chonnam National University, Gwangju, 61186 Korea
[2]Department of Electronic Engineering, Mokpo National University, Jeonnam, 58554 Korea
[3]Rural Development Administration, 300, Nongsaengmyeong-ro, Jeonju-si, South Korea

{whcho, nmh}@chonnam.ac.kr, narciss76@mokpo.ac.kr, leehr26@korea.kr

Abstract. Generally, the occurrence of insect pests in the cultivation of facility or field vegetables is greatly influencing their yields. Therefore, we are very important to accurately detect pests during vegetable life. In this paper, we have considered several deep learning algorithms that can detect various insect pests caused by tomato on the basis of the large - scale pest damage data observed in the facility or field. Here, the deep running algorithm considered is Faster R-CNN with ssd_mobilenet v1 and inception v2. We have considered 4 kinds of pests, which are frequent in tomatoes. We also used 850 images of DSLR images of these pests. From the experimental results, we found that the object detection rate of inception v2 is 50% and that of ssd_mobilenet v1 is about 55%. Therefore, inception v2 showed better performance than ssd_mobilenet v1 in detection of insect pests.

Keywords: Automatic detection of insect pests, Tomato diseases, Protection of diseases, Deep learning algorithm, Faster R-CNN with Inception v2 and ssd_mobilenet v1

1 Introduction

The pests of vegetables can cause big damages to agriculture crops which decrease the production significantly. Early blight is a typical example of diseases that can severely decrease the production. Similarly, in a humid climate, late blight is another very destructive disease that is able to affect the plant leaves, stems, and fruits. Protecting vegetables from various diseases is vital to guarantee crops quality and quantity. Successful strategy of protection should start by an early detection of the disease in order to choose the appropriate treatment at the right time to prevent its spreading.

Usually, this detection is achieved by farmers trained by a practical experience on diseases symptoms and causes. Furthermore, these farmers must monitor vegetables consistently to avoid disease spreading. This continuous monitoring represents a difficult and time-consuming task for humans, which makes to need the automation of the vegetables diseases detection and identification essential to protect vegetable diseases.

Recently, several studies have suggested using image processing and machine learning to detect and classify vegetative diseases. This approach attempts to create a disease classifier using images from plants. These classification methods extract relevant information for image classification based on the manual functions designed by existing experts. For this reason, these classifications lack automation due to the manual dependency of functionality.

To solve these problems, in recent years, deep learning (DL) algorithms have been proposed that surpass advanced technologies in many fields. The main advantage of Deep Learning in computer vision is the direct exploitation of image without any hand-crafted features. Deep Learning classifiers are end-to-end systems that form features in a fully automated way without any interference of human experts.

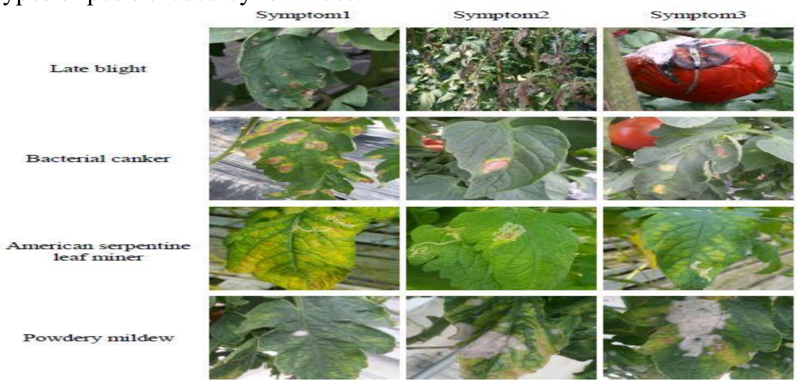
Therefore, in this paper, various pesticide images taken from tomatoes grown in greenhouses were used as database, and to classify these pathogens and identify the pathogenesis, we used the Faster RCNN method, one of the recently developed deep-learning methods.

2 Dataset and Method

In this section, let us consider the collection process of the data set used for the insect pest discrimination and the deep learning algorithm used to determine the insect pest using the collected data.

2.1 Dataset preparation

We have prepared the learning data to detect tomato pests using the deep learning algorithm in the following steps. The first step is the data collection phase. This is an important stage for developing any data-driven application. We collected about 2,000 various image dataset using cameras directly the 4 types of insects occurred in tomatoes grown on the green houses. Figure 1 below shows representative images of the 4 types of pests caused by tomatoes.



.

Fig. 1. Four types pests symptoms of tomatoes

The second step is to label the collected images. To do this, we selected 200 images with comparative clarity and discrimination power for each of the 4 types of insects from the collected images. We used totally 800 images for the experiment. The labeling process consists of annotating the collected images by a human expert. Here, the agriculture expert identifies only the disease in each tomato without any additional information about this disease. The third step is to generate annotation XML for the area where the insect pests have occurred in each image through labeling.

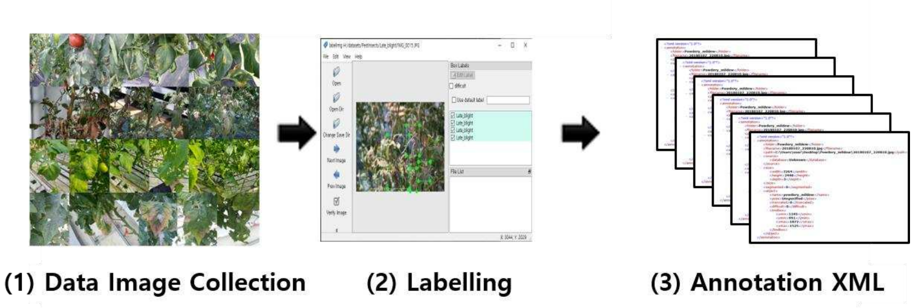


Fig. 2. Preparation process of learning data

2.2 Method

The deep learning algorithm used in this experiment is a Faster RCNN model with the fastest processing time and the best recognition rate. Figure 3 below shows the approximate structure of the Faster RCNN.

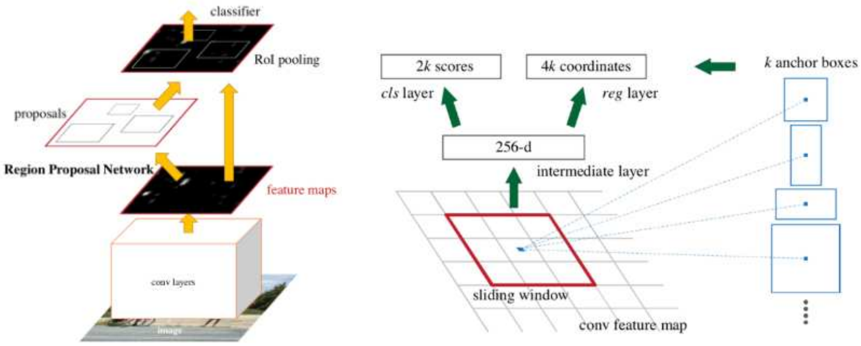


Fig. 3. Approximate structure of Faster RCNN.

The image is provided as an input to a convolutional network which provides a convolutional feature map. Instead of using selective search algorithm on the feature map to identify the region proposals, a separate network is used to predict the region proposals. The predicted region proposals are then reshaped using a RoI pooling layer which is then used to classify the image within the proposed region and predict the offset values for the bounding boxes.

3 Experimental Results

3.1 Training Deep Neural Networks

In this experiment, we consider two state of the art deep neural networks (Faster R-CNN with MobileNet and Inception v2). To learn and evaluate the performance of these state of art CNN, we used the Python Deep Learning framework called Tensorflow with a GPU acceleration option.

All of deep network models are training based on two stages which are Pre-training and Fine-training. First, Pre-training consists of training a deep CNN on a large dataset like ImageNet first, before the training on our dataset. This pre-training is achieved in order to prepare the CNN by the transfer learning from a big dataset to plant diseases detection and identification. This stage is used to deal with the lack of labeled data in plant detection and identification. Second, Fine-tuning is that the last layer (output layer) of the original pre-trained network is replaced with a new layer compatible with the number of classes in our dataset. The obtained network is then retrained using the backpropagation algorithm to fit our dataset. This method improves the results of our model because the weights have already been trained on a bigger dataset. This fine-tuning is a transfer learning way that allows plant diseases task to take advantage of models trained on another computer vision task where a large number of labeled images is available. Finally, we used the commonly used LabelImg to get the labeled image.

3.2 Model Evaluation Results

The dataset is randomly divided into 80% for training and 20% for evaluation. All experiments are performed on a powerful machine, having the specifications that are summarized in Table 1.

Table 1 Machine characters

No	Hardware and software	Characteristics
1	Memory	16 Gb
2	Processor (CPU)	Intel Core i7-4790 CPU 3.6 GHz x8
3	Graphics (GPU)	GeForce GTX TITAN X 12Gb
4	Operating system	Linux Ubuntu 16.04 64 bits

The accuracy for two CNN for detection and identification of tomato diseases is displayed in Table 2.

Table 2 Confusion matrix for two deep learning algorithms

(a) faster R-CNN mobilenet

	1	2	3	4	5	total
1	40	1	0	0	0	97.6%
2	0	7	0	0	28	20.0%
3	0	0	13	1	21	0.37
4	0	0	0	23	15	39.5%

(b) faster R-CNN inception

	1	2	3	4	5	Total
1	27	3	5	3	1	69.2%
2	6	14	6	6	6	36.8%
3	5	3	19	1	4	59.4%
4	5	3	5	29	0	69.0%

※ 1: Bacterial canker, 2: Late blight, 3: Leaf miner, 4: Powdery mildew, 5:Not detection

From results of Table 2, we observe that the most successful learning strategy in vegetable diseases detection and identification for all CNN architectures is the deep transfer learning. Also, we can observe that Faster R-CNN inception has been shown to achieve a classification rate of 60%, and Faster R-CNN mobilenet is found to achieve 55%. Therefore, Inception net seems to have better classification rate than Mobilenet. In addition, the reason that both deep learning algorithms are generally low in the classification rate is that there is not enough learning data and that farmers shoot arbitrarily when they take images in the field.

4 Conclusion

In this paper, we considered the problem of detecting and identifying the pests of tomatoes by Faster RCNN model and two deep learning algorithms like as Inception.v2 and Mobilenet. We have considered 4 kinds of pests (Bacterial canker, Late blight, Leaf miner, Powdery mildew), which are frequent in tomatoes. We also used 850 images of DSLR images of these pests. From the experimental results, we found that the object detection rate of inception v2 is 95% and that of ssd_mobilenet v1 is about 90%. Therefore, inception v2 showed better performance in detection of insect pests by ssd_mobilenet v1.

Acknowledgments. This work was partially supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET), funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (319003-01) and the Korea National Research Foundation (Project No. 2017R1D1A1B03028808) of Korea Grant funded by the Korean Government.

References

1. Barbedo, J. G. A., A review on the main challenges in automatic plant disease identification based on visible range images, Biosystems Engineering, Volume 144, April 2016, Pages 52-60.
2. Golhani, K., Balasundram, S. K., Vadamalai, G., Pradhan, B., A review of neural networks in plant disease detection using hyperspectral data, Information Processing in Agriculture, vol. 5, pp 354-371, 2018. (2006)
3. Brahimi, M., Arsenovic, M., Laraba, S., Sladojevic, S., Boukhalfa, K., Moussaoui, A., Deep Learning for Plant Diseases: Detection and Saliency map Visualization, Human and Machine Learning pp 93-117.

Determination of vegetables grade using pattern recognition method and deep learning algorithm

Jun ki Kim[1] , Wanhyun Cho[1], Myunghwan Na[1], Sangkyoon Kim[2], Hye Jin Lee[3]
[1] Dept. of Statistics, Chonnam National University, Gwangju, 61186 Korea
[2]Dept. of Electronic Engineering, Mokpo National University, Jeonnam, 58554 Korea
[3]Rural Development Administration, 300, Nongsaengmyeong-ro, Jeonju-si, South Korea

kjkwnsrl@gmail.com,{whcho, nmh}@chonnam.ac.kr, narciss76@mokpo.ac.kr,
lhj5157@korea.kr

Abstract. In this paper, we mainly focus on the prediction of grading algorithm of vegetable, which are essential factors for sorting and grading agricultural products. First, we investigate the color models and several features that can be used to properly represent the freshness, shape and quality of agricultural products. Second, we consider pattern recognition that can be used to assess the quality of agricultural products and to determine their grades. Third, we identify whether the combination of feature vectors and classification methods is the most efficient in evaluation the quality and the grading of agricultural products through experiments on various vegetable images. From the experimental results, we found that VGGNet16 shows a slightly better classification rate than SVM with HOG. These results are expected to be useful for automatic grading of crops.

Keywords: computer vision system, grading and sorting of vegetables, SVM with HOG feature, deep learning algorithm vggnet16, classification rate

1 Introduction

Recently there has been a growing interest in the use of computer vision technology in technology in agriculture. Typical areas of using computer vision technology in agriculture include crop monitoring, precision agriculture, cultivation using robotics, automatic guidance, non-destructive inspection of product properties, quality control and classification on processing lines and process automatic control. As mentioned above, computer vision technology is used in many fields because it provides potential information about the various characteristics of crops that cannot be detected by the human eye.

In this paper, we mainly focus on color feature extraction methods and classification algorithms, which are essential factors for sorting and grading agricultural products such as fruits and vegetables. First, we investigate the color models and color features that can be used to properly represent the freshness and maturation stages of agricultural products. Second, we consider various artificial intelligence classification algorithms that can be used to assess the quality of agricultural products and to determine their grades. Third, we identify whether the combination of color feature vectors and classification methods is the most efficient in maximizing the classification rate in the grading of agricultural products through experiments. Finally, we conclude the results of the experiments and future research directions.

2 Dataset and Methods

The essential characteristic for food quality evaluation using computer vision technology is to classification which contribute with human beings so that they can easily select good quality products when purchasing agricultural products. To accomplish this task, vegetables images can be described by set of features such as color, size, shape and texture by using image processing techniques. Here, we consider the dataset of cucumber from collected by Gyeonggi-do agricultural Research & extension Services and two machine learning algorithms which are both Support Vector Machine (SVM) and Convolution Neural Network (CNN).

2.1 Dataset

In this paper, we use the dataset of cucumber from collected by Gyeonggi-do agricultural Research & extension Services. The data for evaluating the VGGNet model was created by Gyeonggi Provincial Agricultural Research Institute with a total of 100 cucumber image data. The characteristics such as shape, length, size, degree of warpage, color, texture, scratching, and thorns were influenced by the grades. Therefore, the grades of cucumbers were classified as high, middle, low products. To prevent the lack of a dataset for deep learning, we made the total dataset 400 by flip and 180 degree rotation of the image using data augmentation technique. Of these, 280 were used as training data and 120 were used as test data. In addition, in order to remove unnecessary background, the data was adjusted to 150~400 width and 350~880 length.

2.2 Methods

In this study, we were attempted investigate the best choice among SVM kernels for speaker identification task. Before that, HOG provides the necessary features. HOG makes the shape of the object in the image through the edge direction distribution of the cucumber. The image is divided into small connection areas, in which a bar graph in the direction of the gradient is compiled. Then, the descriptors are combined by connecting different histograms. The HOG descriptor extracted the features from the color image using the factors of number of orientation bins of 8, size of cell of (16, 16), and number of cells of each block of (1, 1). Fig. 1 shows the process of feature extraction of HoG. We use SVM model with 3690 features as inputs.

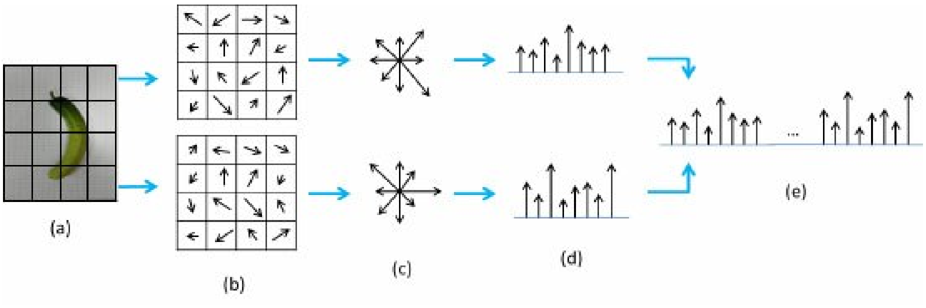


Fig. 1. HOG feature extraction process.

VGGNet16 used in this experiment consists of two convolution filters 64 and one max-pooling, which are used as one convolution layer. They used the first layer with 64, the second 128, the third 128, the fourth 512, and the fifth 512 filters. Finally, we used three Fully Connected layers and the last on for sorting. We used ReLu as an activation function in all layers, Soft-max as a classification, He-uniform as an initialization method, and Cross-entropy as a loss computation. The learning method creates the structure of the model as above and imports the weight storage file that has already been learned, and starts the training using the cucumber image as input data. This can reduce training time and provide better performance. Fig. 2 shows the cucumber image of each grade applied to the SVM HOG descriptor.

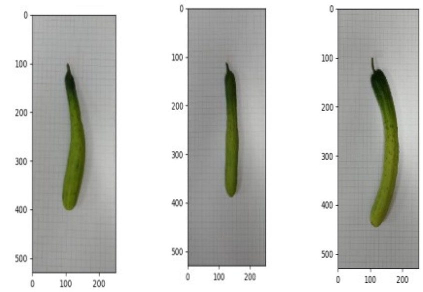


Fig 2. Visualization by each grade.(Low, Middle, High)

3 Experimental Results

Experimental results shows that training time of VGGNet is definitely faster than SVM, which is a traditional machine learning method because it took a lot of time to extract the HOG feature. Moreover, the confusion matrix for test data is given Table 1. It showed SVM 90% of low, 50 % of middle, 50% of high, and the total accuracy was 70.8% as a whole and VGGNet is 90% of low, 52.5% of middle, 70% of high, and the total accuracy was 70.8% as a whole.

Table 1. Classification rates for our model

	Low	Middle	High	Total
Acc_SVM	0.90	0.50	0.50	0.633
Acc_VGGNet	0.90	0.525	0.70	0.708

4 Conclusion

Here, we examined traditional machine learning methods and Deep learning algorithms which are essential in computer vision systems that select and grade agricultural produce such as vegetables.

Experiments were carried out to classify cucumbers using SVM models using HOG feature extraction models and VGGNet based on CNN. A comparison of the two model showed that VGGNet was more accurate and faster. We expect that if we collect more learning data, we can further improve of the classification because the result of the VGGNet classification is that the less accurate reason is the lack of learning data.

In the future, the study considers the issue of how to choose appropriate other characteristics for grading, and is also studying classification methods to determine the freshness of vegetables.

Acknowledgments. This work was partially supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET), funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (319003-01) and the Korea National Research Foundation (Project No. 2017R1D1A1B03028808) of Korea Grant funded by the Korean Government.

References

1. McConnell, R.K., Method of and apparatus for pattern recognition, United States: N. P., 1986
2. N. Dalal, B. Triggs., Histograms of Oriented Gradients for Human Detection., International Conference on Computer Vision & Pattern Recognition (CVPR '05), Jun 2005, San Diego, United States. pp.886--893,
3. Vapnik, V., Lerner, A. (1963). Pattern recognition using generalized portrait method. Automation and Remote Control, 24, 774-780.
4. Alex, K., Ilya, S., Geoffrey E. H. (2012). Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, 1097-1105.
5. Kwon, S., Cho, H. (2014). A Simple Graphical Method for Detecting Consistently Misclassified Samples Using Robust SVM, Journal of the Korean Data Analysis Society, 16(1), 125-133. (in Korean)

6. Karen, S., Andrew, Z. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR2015.
7. Keras: The Python Deep Learning library, https://keras.io/
8. Scikit-learn Machine Learning in Python, https://scikit-learn.org/stable/

Extraction of Impact Factors Related to Onion Disease Using LDA

Jin-Hyun Song[1] , Tserenpurev Chuluunsaikhan[1], Jeong-Hun Kim[1],
Kwan-Hee Yoo[1], Hyung-Chul Rah[2], Aziz Nasridinov[1,*] ,

[1]Department of Computer Science, Chungbuk National University, South Korea
[2]Big Data Cooperative Course, Chungbuk National University, South Korea
thdwlsgus10@chungbuk.ac.kr, teo@chungbuk.ac.kr, etyanue@chungbuk.ac.kr,
khyoo@chungbuk.ac.kr, hrah@cbnu.ac.kr, aziz@chungbuk.ac.kr
(*Corresponding Author)

Abstract. Recently, various cooking TV shows have increased the supply and demand for onions production. Onion has many ingredients for a healthy lifestyle. However, concerns related to the onion disease are regularly occurring, and impact factors for disease are not easily understood. Previously, studies have been carried out on improving onion's taste and other ingredients. However, studies on the diseases affecting the quality of onion are insufficient. In this paper, we propose a method for extraction of impact factors related to onion disease. For this, we first collect news articles related to the onion. Next, we analyze the collected articles using the LDA algorithm and extract impact factors related to the onion disease. Lastly, we propose a web interface to show impact factors to the users. We expect that the proposed method can help to understand better regarding the onion disease.

Keywords: Onion disease, LDA, Data collection, Data visualization

1 Introduction

The quality of onions is directly related to the demand of consumers, and if the quality is good, it can have a positive effect on demand. However, concerns related to the onion disease are occurring; this affects the quality of the onion. Recently, studies on improving onion's taste and other ingredients have been carried out a lot. However, studies on the diseases affecting the quality of onion have not been carried out correctly. Therefore, the study is needed to analyze the factors affecting the disease.

In this paper, we propose using LDA algorithm to find the impact factors of disease affecting onion quality. To do this, we first collect news articles and remove unnecessary words and phrases from the collected articles. Further, the LDA algorithm is used to analyze the impact factors through the distribution of words per topic. Finally, we propose a web interface that visualizes the result obtained from the analyzed data. More precisely, we make the following contributions in this paper.

- We propose a method that collects and processes large amounts of data. The proposed method removes unnecessary words or phrases from many news articles and prepares only necessary nouns.
- We propose to analyze the nouns using well-known LDA algorithm to extract impact factors related to onion disease such as onion nocicept, and leaf blight.
- We propose a web interface to visualize the result of our analysis. User can identify the impact factors related to onion disease by year using the web interface.

This paper proceeds as follows. In section 2, we briefly introduce the other studies using LDA algorithm. In section 3, we introduce the proposed method. In section 4, we explain the experimental data, and the result of experiment. Finally, in section 5, we describe the conclusion of the study.

2 Related Study

Text mining is being carried out to analyze based on text information such as newspapers, online news articles, and blogs. In-text mining, topic modeling method is in progress to find out the distribution of topics in each document and the distribution of topics by year [1]. One of the topic modeling methods, the Latent Dirichlet Allocation (LDA) is a type of unsupervised learning. Each document is a set of topics, and each topic is a set of words. By classifying them based on topics that are latent in the document, it can be used when there are several topics in the document [2].

The study [3] analyzes the core issues and research trends of national policy by analyzing Korean news articles using LDA algorithm, and word cloud. It predicted renewable energy with potential for growth. As a result, it found 20 topics related to the renewable energy field. Suh and Shin [4] analyzed the academic papers on the platform government of dbpia, which is a national dissertation database, using R-based topic modeling packages such as Tm, Ldavis, and Stm, and LDA algorithm based on Platform - Government Research Trends. As a result, it could derive each topic according to the words distributed in the topic related to platform government.

Many other studies have also derived important topics using topic modeling methods. Likewise, this study uses the topic modeling method to identify impact factors related to onion disease by using only news articles related to onion disease.

3 Proposed Method

We propose a method of extracting impact factors through news articles on onion disease. The method of this study proceeds shown as in Fig. 1. First, Data Collection and Preprocessing module collects news articles related to onion disease and preprocesses it using the procedures of Konlpy morpheme analysis. Second, the LDA topic modeling module extracted the main representative words of the articles. Third, Data Visualization module consists of three graph: Line Chart, Word Cloud, and Pyldavis to show the extracted words.

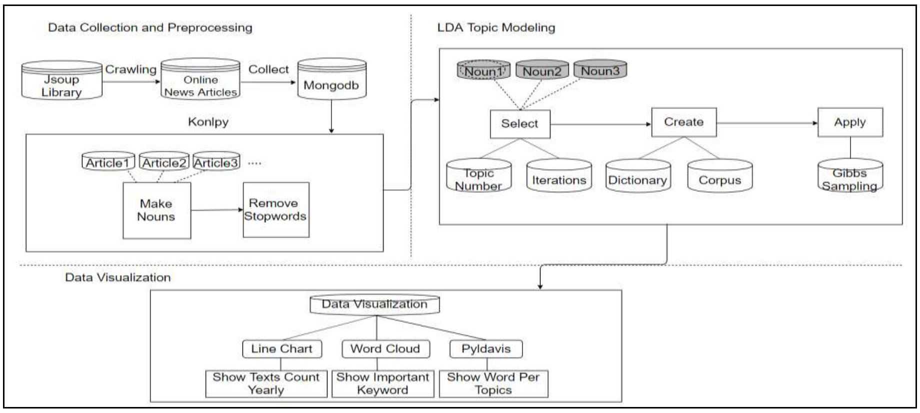


Fig. 1. The overall structure of extracting impact factors related to onion disease

3.1 Data Collection and Preprocessing Module

We collected news articles related to onion nocicept and leaf blight, which is frequently mentioned in the news articles among related to onion disease from 2010 to 2019. News articles were collected using Jsoup, a java-based on Data Collection Library, and Mongodb, which is a non-relational database. The contents of the document which includes Search word named the onion and content called the onion nocicept and leaf blight in the body of the text. In the Preprocessing step, we tokenized articles, which change to noun words using python library Konlpy, and we removed unnecessary words.

3.2 LDA Topic Modeling Module

After extract nouns, we use LDA topic modeling. LDA topic modeling module has three procedures. First, The LDA topic modeling to be used in this study should specify the number of the topic to be extracted and the number of iterations to perform the modeling. Second, prepare a collection of words, a dictionary, and a corpus to manage indexes and words together. Finally, the method to decide the parameter uses the Gibbs Sampling algorithm. Gibbs Sampling is a process of sampling. When there are two or more variables, we are sampling one by changing one variable and then continuously changing another variable. In this study, Gibbs Sampling was repeated 1000 times to extract three topics. The subject name was selected through extracted topics.

3.3 Data Visualization Module

To visualize the result of the LDA topic modeling, we use the d3.js library for using the graphs: Line Chart, Word Cloud, Pyldavis. The purpose of the Line Chart is to visually show how much of the year related to onion disease text is. A Word Cloud is a technique used to express words and topics visually. Pyldavis is a library to visualize the result of LDA topic modeling. Through Pyldavis, we can predict the topic through a combination of words distributed for each topic.

4 Performance Evaluation

In this section, we describe data used in this experiment and explain the result analysis.

4.1 Experimental Data

In this subsection, we describe the data used in the experiment. The data used in this experiment is divided into three procedures, Data Collection step. Before Preprocessing step and After Preprocessing step. Data Collection step has consisted of unprocessed news articles were selected from the portal 'N.' This data has the search term of onion disease related to articles from 2010 to 2019, and words containing the disease words in the text contents.

Before Preprocessing step is consisted of unprocessed handling stopwords and tokenized nouns. After Preprocessing step is comprised of processed handling stopwords and tokenized nouns. Data situation per step is as shown in Table 1.

Table 1. Data situation per step

Steps	**Data number**
Data Collection	181 texts
Before Preprocessing	8461 words
After Preprocessing	1061 words

4.2 Result of Experiments

In this subsection, we describe the experiment result using three graphs: Line Chart, Word Cloud, and Pyldavis. We introduce the Line Chart graph. We can see the Line Chart graph shown as in Fig. 2. The x-axis of the graph is the year from 2010 to 2019, and the y-axis of the graph is the number of news articles by year.

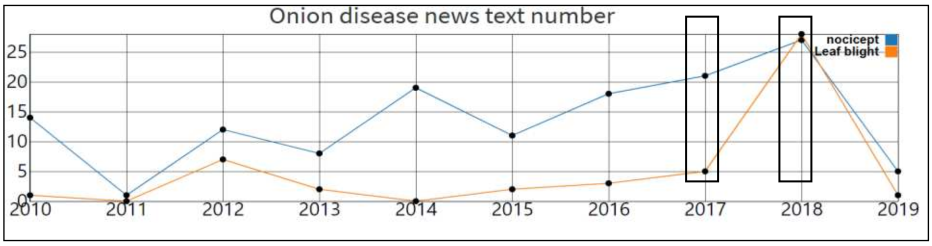


Fig. 2. Onion disease news articles number

We looked at the line graphs and found that the two points where the number is high are 2017 and 2018. We have drawn a Word Cloud of news articles in 2018, which has the most news articles, as shown in Fig. 3 and we compared the words that occurred in the other year, as shown in Fig. 4. In Fig. 3, we could guess that the onion nocicept and leaf blight affected the production of onion through the words of Damage, Farmers, Farms, and Production.

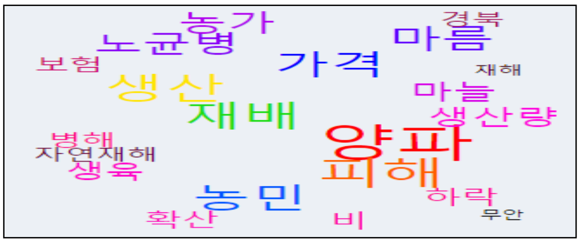


Fig. 3. Word cloud for 2018 news articles

We can see words related to onion disease from 2010 to 2019. We could find some common words, through a combination of common words we saw, we considered factors that affect onion disease. First, in 2012 and 2014, words such as Price and Nocicept were found. It was found that onion disease affects the Price. Second, in 2010, 2012 and 2019, a word such as Prevent was found. It found that people are paying attention to the prevention of onion disease. Finally, in 2010, 2011 and 2012, the words of Rain, Drought were found. It found that it was related to disaster and climate.

Table. 2. word distribution by year

Year	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	Top9	Top10
2010	Onion	Nocicept	Damage	Plantation	Rain	Farm	Growth	Harvest	Prevent	Soil
2011	Rain	Garlic	Farm	Temperature	Damage	Onion	Sow	Disaster	Crop	Farming
2012	Onion	Price	Plantation	Nocicept	Garlic	Drought	Pesticide	Damage	Prevent	Harvest
2013	Onion	Nocicept	Plantation	Pest	Farm	Muangun	Damage	Income	Soil	Infection
2014	Onion	Price	Nocicept	Garlic	Plantation	Farm	Output	Kyungbuk	Produce	Excess
2015	Onion	Garlic	Nocicept	Plantation	Prevent	Farm	Price	Damage	Produce	Output
2016	Onion	Nocicept	Garlic	Plantation	Price	Prevent	Rain	Damage	Release	Cabbage
2017	Onion	Nocicept	Prevent	Drugs	Garlic	Pest	Farm	Plantation	Rain	Damage
2018	Onion	Damage	Produce	Plantation	Farmer	Price	Dry	Nocicept	Farm	Garlic
2019	Onion	Nocicept	Prevent	Infection	Garlic	Dry	Rain	Damage	Temperature	Mold

To see the topic per words, we use Pyldavis using d3.js. We experimented with setting the number of iterations to 1000 for three topics. In Fig. 5, we can see that three topics have been created. In the essential topic, words such as Drought, Rainfall, Disaster, Damage, Pests indicate that this topic is the climate one of the onion disease impact factors.

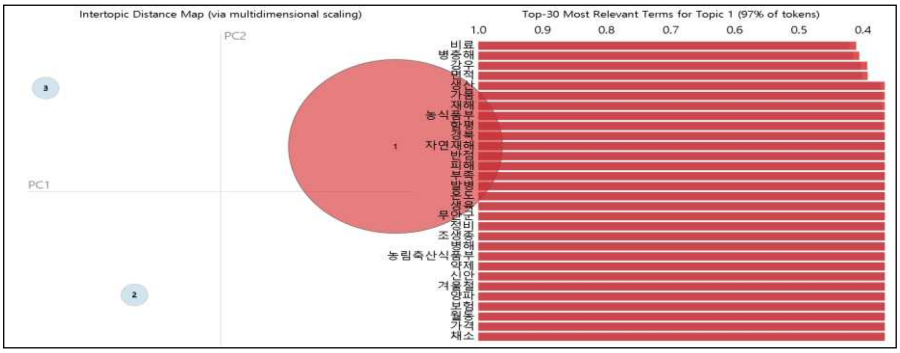


Fig. 5. Topic extraction for using Pyldavis

5 Conclusion

In this paper, we extract the impact factors by using the news articles of onion nocicept and leaf blight among the onion related to disease and analyzed the impact factors through the extracted words. The result showed that seasonal factors influenced onion nocicept and leaf blight because of the onion cultivation from the outside. We could see words such as Price and output per year. This shows that the Price and yield of onion is one of the essential factors affecting the news articles. Besides, the words of garlic in 2011 and 2015 are also referring to the cultivation of garlic is also very relevant. The improvement afterward is that the distribution of the words is not consistent for each topic. So it is necessary to experiment with the texts of various the topics and the news articles of more kinds of disease.

References

1. No Byeong joon., Xu Zhenshun., Lee Jonguk., Park Daihee., Chung Yonghwa.: Analysis of foot-and-mouth disease spreading effect based on keyword network using online news, Journal of KIIT. Vol. 14, No. 9, pp. 143-152, Sep. 30 (2016)
2. David, M, Blei., Andrew, Y, Ng., Michael, I, Jordan.: Latent Dirichlet allocation, The Journal of Machine Learning Research, Vol. 3, pp. 993-1022, Jan (2003)
3. KyuSik Shin., HoeRyeon Choi., HongChul Lee.: Topic Model Analysis of Research Trend on renewable energy, Journal of the Korea Academia-Industrial cooperation Society. Vol. 16, No. 9 pp. 6411-6418 (2015)
4. Suh Byung-Jo., Shin Sun-Young.: A Study on the research on domestic platform government using topic modeling, Information Technology Policy Vol 24, Issue 3, pp.3-26 (2017)

Acknowledgments.

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET), funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (319003-01)

Eco-friendly Irrigation System Using LoRa

Jeongyeon Yu[1], Jiae Yun[2], Hyunmuk Kim[3], Jordan Kramer[4], Minsun Lee[3]

[1] Deptartment Of Environmental Engineering, [2] Department of Statics, [3] Division of Computer Convergence,
Chungnam National University, Daejeon 34134, Korea
[4] Deptartment Of Computer and Information Technology, Purdue University, USA
{dbwjddus1997, yja2397, eszesz321}@gmail.com, kramer60@purdue.edu,
mleeoh@cnu.ac.kr

Abstract. Due to climate change, drought is becoming a more prominent global issue, and there is a need to manage limited water resources effectively. Traditional irrigation systems do not monitor the water content of soil and apply the same amount of water to wet areas as dry areas. In this paper, we propose a web integrated and smart irrigation system using a LoRa wireless communication. We demonstrated the proposed system based on sensor data and weather information is eco-friendly farming methods by regulating the amount of water.

Keywords: LoRa, Wireless communication, Smart farm, Irrigation system

1 Introduction

In this century on a global scale, drought is a major environmental problem. Since the 2000s, droughts have frequently occurred in many parts of the world. Some of the leading causes of drought are reduced rainfall, shorter rainy seasons, and less typhoon activity. There has also been less water supply because of outdated facilities. Pipes, reservoirs, and water facilities are old and developing leaks leading to the reduction of water [1, 2].

The purpose of this research is to analyze the ways to reduce water waste and incorporate findings on current automatic irrigation systems, soil characteristics, current scalable water systems as well as information about crop yields. We accumulate sensor data about humidity, temperature, and transmission methods, and design an automatic irrigation system using the LoRa gateway. We build a “Wayne’s Crop” website that integrates with the proposed system of checking real-time conditions and immediately controlling farm emergencies. Our proposed system uses humidity and temperature sensors to read the soil condition and then uses web applications to control water solenoid valves to supply water. In this study, we consider corn, one of the most widely cultivated crops in the world. The experimental site is located in Indianapolis, Indiana, USA.

2 System Design and Implementation

2.1 LoRa Communication

LoRaWAN(Long-Range Wide Area Network) can transmit small amounts of data over long ranges. It supports bi-directional communication between the gateway and end-nodes [3]. LoRa device signal can reach 8km in urban settings; however, the signal strength is much greater in rural areas because of fewer obstructions [4]. We used LoRa application technology in our experiment environment. Similar to smart irrigation or automatic control systems, the LoRa systems are more economical because they can cover wider areas and consume less electricity.

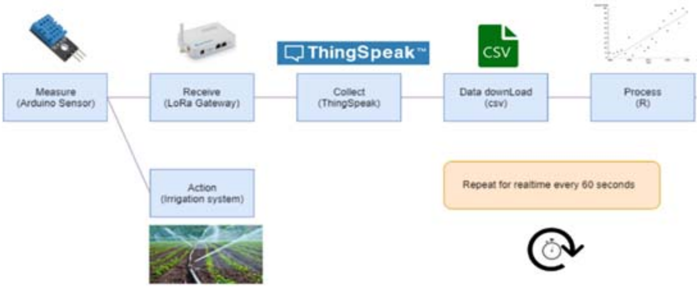


Fig. 1. Proposed System Diagram

We used a node using the Arduino UNO board and LoRa shield with two sensors. Three measurements were transmitted wirelessly to the gateway, LG01. The solenoid valve for irrigation was controlled by the relay module. A VH400 Soil Moisture Sensor was used to check soil humidity. It uses superior transmission line techniques (TDR) to measure the water moisture in any soil regardless of soil salinity. Probes do not use exposed metal, so they never corrode or need to be recalibrated [5]. Another sensor is DHT11, for air humidity and temperature. DHT11 includes two types of parts. One is the NTC (Negative Temperature Coefficient) temperature measurement component, and the other is a resistive humidity measurement component

LoRa devices are composed with a gateway and LoRa shield. Gateway LG01 runs on Arduino code. It provides a wireless network bridge to IP networks based on WiFi, Ethernet, 3G or 4G cellular. We used this Gateway to connect the node to ThingSpeak [6]. The frequency range is 862-1020 Mhz for HF and 410-528Mhz for LF.

2.2 Software

Arduino software (IDE) is the main controller of this system. We use the Bridge library, the Process Library and the HTTP Client library to control LoRa communication systems. The Bridge Library defines a mechanism for how the MCU talks to the CPU. It works for bi-directional communication. The MCU can send data, and receive commands from the CPU [7]. The Process Library is used to transmit sensor data and control valve opening. The process is the basic class of Bridge-based calls, which allows the gateway to send data to ThingSpeak server and website. The solenoid valve can be controlled by our web application. When the valve setting is switched to “on”, the HTTP Client library obtains the control information by calling the “get” function and changing the flag to 0 (off) or 1 (on).

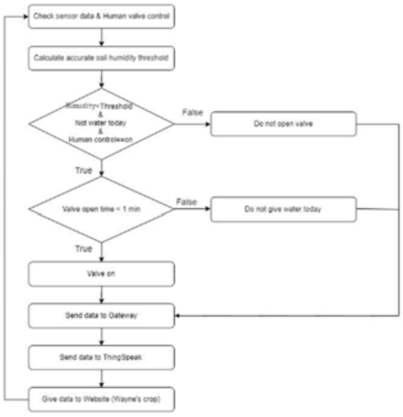


Fig. 2 Workflow of the proposed system

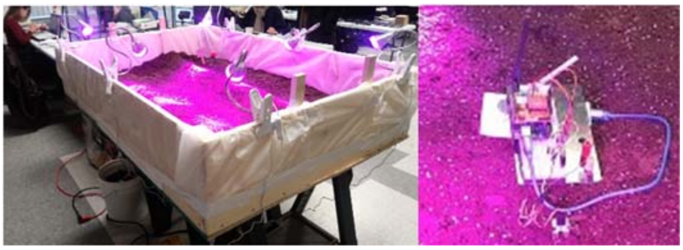


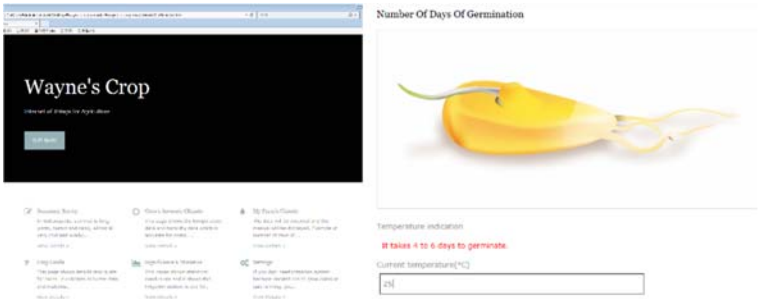
Fig. 3. Cedar box farm and Sensors

2.3 Cedar Box Farm

As shown in Fig. 3, a small-sized (152.4cm(L) x 101.6cm(W) x 30.48cm(H)) box farm was constructed. The water supply has sufficient hydraulic pressure to flow. If the humidity is below the optimal level, the water pump irrigated our plants for one minute. After one minute of operation, the pump stopped for 24 hours to allow the soil to absorb water. This experiment was carried out to confirm the efficiency of using an automatic irrigation system.

3 Results

We built a website named “Wayne’s Crop” that showed sensor data and analyzed the results (Fig.4). The website consists of seven subpages. The “Summary Survey” page shows Indianapolis' climate and information from the LoRa. The sensor provides temperature and humidity data and communicates through the LoRa gateway. It updates the data every 15 seconds, as it appears on the "My Farm's Climate" page. We then compared this data to the optimal temperature and humidity levels of corn cultivation obtained from the official site of Rural Development Administration[1] and display this on our "Corn's Accurate Climate" page. The "Crop Guide" page displays an accurate number for days of germination, growth temperature and so on (Fig.4). The "Significance & Statistics" page displays all systems, including irrigation and standard systems. It also displays the water savings of our irrigation system. For example, if irrigation is not needed in the event of harvest or heavy rain, the system may be deactivated on the "Settings" page. This increases the efficiency of farming and can use less water. The "Gallery" page has our experiment processes and the results displayed with videos and pictures.



Wayne's Crop

Number Of Days Of Germination

Temperature indication

It takes 4 to 6 days to germinate.

Current temperature(°C)

25

Fig. 4. Main page (left) and Crop Guide page (right) of the Website

If a user inserts a temperature value into variable x through the website, the user can obtain a humidity value appropriate for that temperature. The results are graphed with the actual data shown in Fig. 5. The black dot is the current humidity and the blue dot represents the appropriate humidity. The irrigation system operates when humidity drops below the proper humidity levels observed in the graph.

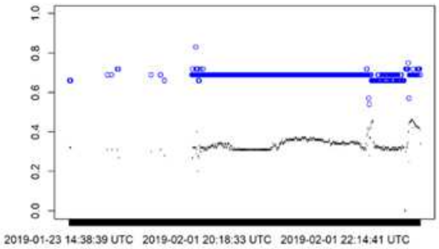


Fig. 5. Comparisons of humidity obtained from Jan. 23 to Feb. 1, 2019.

We designed and made an irrigation system with our sensor and survey data. We compared irrigation data with optimal levels of moisture and climate data to see if there

[1] Rural Development Administration, http://www.nongsaro.go.kr

was any noticeable improvement in crop irrigation using our system. Standard systems do not monitor the moisture content of soil and applies the same amount of water to dry areas as wet areas.

We analyzed nine years (2010 - 2018) of climate data in Indianapolis from June to August. We compared three conditions in Table 1, the first column is climate data: the amount of rainfall in Indianapolis. The second column is the result of hand irrigation, when people give water for corn. The value obtained by adding 240mm to the climate results. The data in the third column is auto irrigation's water consumption, which is less than the manual method. We performed a one-way ANOVA design to confirm that it was statistically significant. Since the p-value is 0.0005407, less than 0.05, so we can reject a hypothesis of "The three amount of water in all three treatments is equal." If we supply water manually and use an irrigation system, we can see that the irrigation system statistically reduces water waste. It means that the proposed irrigation system can save agricultural water.

Table 1. Water amount (Unit: mm)

Year	Climate	Standard	Irrigation
2010	329.438	569.438	482.142
2011	186.436	426.436	386.304
2012	188.722	428.722	410.354
2013	205.740	445.740	300.000
2014	332.486	572.486	368.816
2015	585.470	825.470	646.100
2016	444.754	684.754	444.754
2017	367.792	607.792	422.834
2018	286.258	526.258	323.712

4 Conclusions

In this paper, we presented a web integrated and smart irrigation system based on sensor data and weather information. Environmental condition was measured using LoRa radio connection as a communication standard, and its value was processed by statistical analysis. The system showed water-saving and eco-friendly farming methods by regulating the amount of water.

By expanding this study and applying it to the actual site, the system can efficiently manage a limited amount of water by maintaining the best conditions. Automatic irrigation systems based on sensor data and weather information are important for preparing for long-term droughts.

Acknowledgment

This research was supported by the Korean MSIT (Ministry of Science and ICT), under the National Program for Excellence in SW (2015-0-00930), supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Ministry of Environment & National Drought information-analysis center: 2013-2018 Sustained Drought Analysis & Assessment Report (2018)
2. Li, S.: Application of the Internet of Things Technology in Precision Agriculture Irrigation Systems, International Conference on Computer Science and Service System, pp1009 – 1013. (2012)
3. Nolan, K., Guibene, W., Kelly, M.: An evaluation of low power wide area network technologies for the Internet of Things, International Wireless Communications and Mobile Computing Conference, Sept. (2016).
4. Zhao, W., Lin, S., Han, J., Xu, R., Hou, L.: Design and Implementation of Smart Irrigation System Based on LoRa, 2017 IEEE Globecom Workshops (GC Wkshps) (2017)
5. Vegetronix: VH400 Low-Cost Soil Moisture Sensors, accessed Feb 19, 2020, https: //veg etronix.com/Products /VH400/'
6. Dragino: LG01 LoRa Gateway User Manual, Document Version: 1.4, Firmware Version: IoT Mesh v4.3.3

Visualization of Association Analysis Results in Smart Management Systems

Je-Seog Myeong, Sokchomrern Ean, Kwan-Hee Yoo*

Dept. Of Computer Science, Chungbuk National Univeristy, South Korea
{jeseogmyeong, chomrern, khyoo}@chungbuk.ac.kr
*Corresponding Author

Abstract. The visualization is becoming the advanced tool for presenting the most complex and comprehensive data in the smart management system. This paper aims to give an insight into how to apply the data mining technique to find out the unrevealed relationships between the main features of the fault analysis report. In addition, the paper highlights the results which come from implementing association rule, which is one of the often-used approaches in finding the correlation between items and display into the Sankey diagram and network graph.

Keywords: Visualization, Association Rule, Sankey, Network Diagram

1 Introduction

Understanding the data is extremely important and helpful not only for the decision makers, but also the general audience. In fact, data visualization has been developed and upgraded from a simple to advanced and comprehensive output. That said, it would be straightforward to see a single graph with reflect to the business operation and the main objective rather than look at the equation and the large amount of original data. Moreover, this seems to be a common problem in projecting the results whether in the graph or table. Another key thing to remember is that how to retrieve the usefulness of data and convert the data definition into a user-friendly and meaningful visualization. To overcome this constraint, the association rule of data mining can be used to provide the possibility in terms of projecting the result precisely.

In this study, we propose to apply the association rule results in order to get the significant data in the form of visualization. Furthermore, the main objective of the paper is to showcase the contributions as bellows:

- We propose a rigorous study for visualizing the huge amount of data in the graph with the help of data mining tool to enable user gets easier to understand the graph.
- We choose the association rule which is one of the best methods in figuring out the insight of data with some specific rules which are needed to consider.

- We investigate the results to indicate how much we accomplish the main goal of making the illustration of the data in the form or advanced visualization in the smart manufacturing firm.

Next, the paper is grouped into five sections. Section 2 provides the overview of the related study. Section 3 highlights the proposed method. Section 4 reveals the experimental results of visualization and association rule approach. Section 5 gives a brief summary of the paper and indicates the future work.

2 Related Study

What we know about visualization is largely based upon empirical studies that investigate how to gain insight into the interpretations of complex data and most important part of data pattern. As a result, the previous studies of the visualization and association rule are described below:

It is quite easy to understand the term of visualization in general. Apart from that, it is needed some efforts to know how to select the right graph to present the data properly. Therefore, to depict the complex data set from the result of association rule, Michael and et all introduce the initial method namely grouped matrix-based visualization which enables to display the data with many rules through the clustering [1]. Moreover, it can be aware of the aforementioned investigations that there are two types of method to visualize data with the association rule such as grouping graph and network graph [2, 3].

What is interesting in the data mining is that the term association rule is often used in many fields for learning the unrevealed relationships of data analysis [4]. Additionally, we can note that the crucial distinguish from the association rule to other methods is that there is no fixed rule that can be used to implement in the future to get the same result as we have done nor there is no direct relationship between variables. One of the best examples is that when one variable occurs the other variable does not affect directly [5]. In [6], the author examines two types of process of applying the association rule. The first type is finding the minimum support which comes from the data frequency in a set. Next, the second step is making the rules that fit to the minimum confidence, and removing those rules that are not complying.

3 Proposed Method

For this study, the visualization and association rule are used to explore the effective data analysis in the smart manufacturing management system. The main advantage of the proposed method is that the high volume of data can be presented in a Sankey diagram and network graph. After collecting the data from fault analysis of smart management system, we can obtain the useful information such as line name, machine name, item, sub item, error and treatment. These features are used to find out the association rule with the minimum support and minimum confident are given by the user. Notably, the support is used to show how often the item appears in the set of data. In addition, the confidence shows how the accuracy of prediction. The idea is to find out which sets of machine fault elements occur frequently and visualize in the graphs.

Importantly, the Apriori algorithm is chosen to get into the frequent item set of transaction records. The algorithm starts with determining the counting each item individually in the data set and then resizing them to a bigger set of items until the set of items fit the enough criteria. Several studies agree that one of the best methods to find the minimum support and confidence through the predefined equation [7].

$$\text{Support(X) = count(X)/N} \tag{1}$$

$$\text{Confidence (X => Y) = support (X, Y) / support(X)} \tag{2}$$

In [7], the equation (1) shows the formula of finding the support of item X with a number of data set (N). Next equation (2) describes how to calculate the confidence between item X to item Y. In doing that, we need to point out the support of X and Y.

4 Experimental Results

In this part, the Sankey diagram and network graph serve as the data visualization of the association rule results. Regarding to the dataset, the data are collected from the manufacturing firm which located in South Korea. The size of the data set is 14903 observations. Furthermore, to conduct the experiment, we use a computer with specification such as CPU Intel® Core™ i7-6700 3.40GHz, RAM 8 GB, Graphics card NVIDIA GeForce 9800 GT, Operating System Windows 10 and Integrated Development Environment IntelliJ.

From the Fig 1, we can see that there are six categories. The first category is the line name in the shop floor. The second category is the machine. Interestingly, one line can consist of many machines. The following categories are item, subitem, error and treatment, respectively. According to the graph shown in Fig. 1, we can note the utmost importance of the flow with respect to the association rule.

Another way to show the association rule result is based on the network graph. Fig. 2 depicts the association relationship between one node to another. The node is represented by a circle and the relationship is denoted by the vector or arrow. Therefore, we can get to know that the size of the node describes the support while the width of the vector or edge gives the information about the confidence.

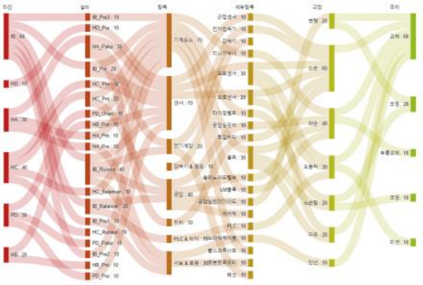


Fig. 1. Visualization of association rule result with the fault analysis

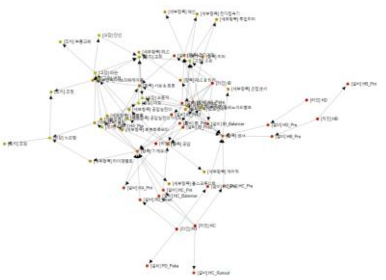


Fig. 2. Network graph of the association rule results with the fault analysis

5 Conclusion

In conclusion, this research extends our knowledge of how to take advantage of visualizing data with the data mining approach to offer some insight into the relationships between set of the data in smart manufacturing management system. All things considered, it seems reasonable to choose the association rule approach to indicate the useful information and present in the form of graphs. However, these findings are limited by the use of only one algorithm without looking closer to other algorithms. Therefore, further work will need to be done in the future.

Acknowledgments. This research was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program, No.10082578, 'Development of intelligent operation system based on big data for production process efficiency and quality optimization in non-ferrous metal industry'.

References

1. Hahsler, M., Chelluboina, S.Z.: Visualizing association rules in hierarchical groups. In 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms. The Interface Foundation of North America (2011)
2. Rainsford, C.P., Roddick, J.F.: Visualisation of temporal interval association rules. In International Conference on Intelligent Data Engineering and Automated Learning, pp. 91-96. Springer, Berlin, Heidelberg (2000)
3. Ertek, G., Ayhan, D.: A framework for visualizing association mining results. In International Symposium on Computer and Information Sciences, pp. 593-602. Springer, Berlin, Heidelberg (2006)
4. Dunham, M.H.: Data mining: Introductory and advanced topics. Pearson Education India (2006)
5. Bach, B.D., Fedja, H., Michael, H.: Evaluation of Position-Constrained Association-Rule-Based Classification for Tree-Structured Data. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 379-391. Springer, Berlin, Heidelberg (2013)
6. Hahsler, M., Sudheer, C.: arulesViz: Visualizing association rules and frequent itemsets. R package version 0.1-5 (2012)
7. Branko, K., Nada, L.: APRIORI-SD: Adapting association rule learning to subgroup discovery. Applied Artificial Intelligence 20, 7, pp. 543-583 (2006)

A CHAID Analysis of Factors Associated with Adolescents' Educational Aspirations: Implications for School Curricula

Jung Park[1], Wan-Sup Cho[2]

[1] Big Data Cooperative Course, Chungbuk National University,
Cheongju, South Korea
[2] Management Information Department, Chungbuk National University,
Cheongju, South Korea
{ refree }@ chungbuk.ac.kr, { wscho }@chungbuk.ac.kr

Abstract. This paper discusses factors that affect the educational aspirations (EAs) of grade 6 students in South Korea. For this, we analyzed data from the Korean Education Longitudinal Study and explored the influencing factors and trends of EAs through the data mining decision tree Chi-square Automatic Interaction Detection (CHAID) analysis; here, results showed that various factors interact hierarchically. Also, we derived the association rules between variables, and with this, suggest the direction for redesigning school curricula for continually improving adolescents' EAs. Overall, this study contributes to the data-driven information pool regarding school curricula designs and highlights the importance of such enhancement with its role in influencing adolescents' future.

Keywords: educational aspirations, school curricula, decision tree, CHAID, Korean Educational Longitudinal Study

[1] Big Data Cooperative Course, Doctor's Program, Chungbuk National University

[2] Department of Management Information System, Chungbuk National University, Professor, correspondent author.

1 Introduction

EAs provide educational horizons or determine the educational climate that would eventually steer students into different social roles, thus becoming indices to students' destinies and future decisions (Deosaran, 1977). EAs have been considered a significant predictor variable of the educational outcomes of students and their choices regarding future careers (Khattab, 2015). However, how the EAs are affected is not due to a single factor but rather a result of the interactions among various factors such as individual difference, home environment, and school life (Berzin, 2010). Cupples et al. (2005) proposed that to explain the concept in which various factors play different roles, one might need to approach it with a data mining strategy rather than explain it through regression analysis. This study explores the factors that may influence the EAs of adolescents (grades 6) in South Korea using the CHAID decision tree. In particular, this study asks the following: (1) How do the determinants of EAs for adolescents, their association with each other? (2) What implications do these results have regarding the school curricula in South Korea? The primary objective of this study is to document factors that influence adolescents' EAs and explore ways to redesign school curricula to help enhance the level of adolescents' EAs in South Korea.

2 Related Research

As EAs focus on planning and acting, it is thus considered as an indicator of the educational level one hopes to achieve: a realistic concept that can affect how students design their future (Khattab, 2015). Students with high EAs can attain various achievements as they are more likely to be proactive and willing to participate in school-related activities, allowing them to have a smoother school life compared to other students. According to Berzin (2010), factors such as gender, family background, institutional context, parent academic involvement, school experiences, and socioeconomic status (SES) are considered to affect EAs during adolescence.

Conceptually, the decision tree is a hierarchical model that consists of a set of rules for the partition of the heterogeneous input data set into groups that are homogenous regarding the dependent variable categories (Milanović & Stamenković, 2016). Decision trees are flowchart-like tree structures that divide the data recursively and display them in a manner similar to an image of a tree (Murphy, 2012). Induction techniques for decision trees include using C4.5, CART, and CHAID techniques. Among these, we will use CHAID. CHAID (Kass, 1980) performs multiway splits using a chi-square test when there is a discrete variable. This method is widely used in the marketing field.

3 Analytical Model and Result

3.1 Analytical Model

In this study, we analyzed the KELS 2013 student survey data from grade 6 in South Korea. Data from 5,553 respondents who responded to the EAs questionnaire were used among 7,324 cohorts. As data mining extracts a combination of variables that explain target variables in a large amount of data as much as it can, it must include many variables that can be used in the analysis (Witten, Frank, Hall, & Pal, 2016). Therefore, it was set to be 38-variables including demographics, cognitive and noncognitive achievements, self-directed learning attitude, participation in class, educational and social psychological environment of school, school life and adaptation, leisure and after-school hours, and parent–child relationship factors. The target variable of this study was EAs, in which case we used the question, “Up to what level of education do you plan on pursuing?” For an effective interpretation of the data, the answers were classified as “Low” (up to high school including middle school), “Ave” (up to university including junior college), and “High” (above graduate school’s course).

R version 3.4.1 was used in this study for preprocessing, and the decision tree was made with SPSS Statistics 24.0 using CHAID algorithm. Furthermore, the maximum depth of the tree was 3, the lowest number of cases of parent node was 100 and was set as 50 respectively for the child node. Bonferroni correction was also used. To eliminate overfitting of the formed model, simple random sampling was implemented to classify the groups as the training (70%) and test datasets (30%).

3.2 Result: Grade 6 decision tree

It was found that class understanding, volunteer activities cognition, academic self-concept, class concentration, creativity, reading preference, self-regulation, career planning, rule compliance, gender, and average hours of studying per day were the factors that affect EAs of grade 6 students <Figure 1>. Among all, class understanding was the main factor that determined EAs of grade 6 students (χ 2=262.718, df=2, p-value=0.000). We also found multiple association rules in this tree. The most impressive association rules are: (1) Node 12 - IF ((Class understanding <= Low) AND (Volunteer activities cognition <= Low) AND (Academic self-concept <= Ave)), then low EA proportion = 24.2%, (2) Node 28 - IF ((Class understanding > Ave) AND (Volunteer activities cognition > Ave) AND (Average hours of studying per day > 2)), then high EA proportion = 48.3%

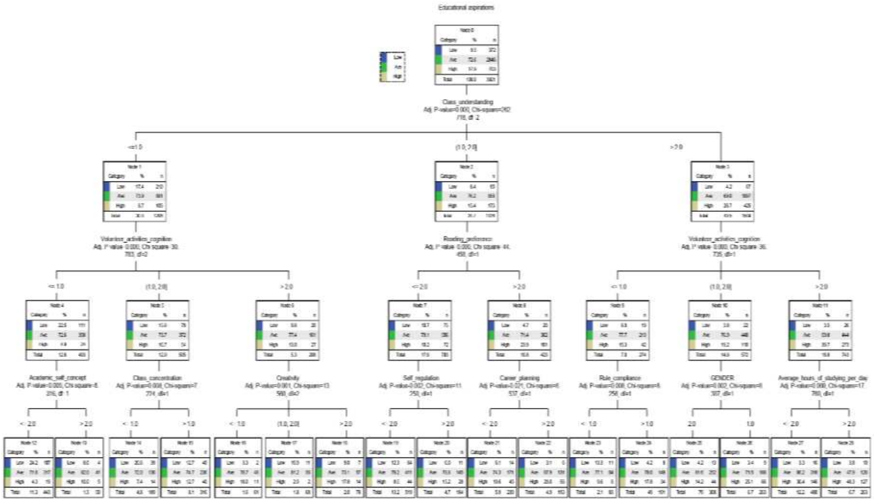


Fig.1 Grade 6 EA decision tree

4 Conclusions and Implications

The results of this study are summarized as follows. We have found that factors are complexly associated to each other about EAs. Class understanding was found as a primary factor, and Volunteer activities cognition was seen as a secondary factor in high and low EAs students. As a third factor, different factors such as Academic self-concept, Average hours of studying per day, etc., were found depending on the level of EAs. Volunteer activities cognition is a secondary factor, which is not easily seen in previous related studies and is seen as a result of data mining methodology.

Based on the above results, we suggest the following direction for the designing of school curricula. First, schools must put the multilateral effort in making studying a habit after school along with minimizing school underachievement. As seen from the study, low class understanding is one of the most important factors that affect low EAs among the students. As such, schools need to work harder on increasing class understanding and plan programs such as setting up of academic clinics that can help students with low class understanding. Such efforts will not only lead an effective school life for students with low EAs but also promote the students' future value. Second, elementary schools must plan and operate educational activities that focus on volunteering for students with low EAs. To improve the student perception toward volunteering, it is important for the students to have hands-on experience.

As this study only inferred its conclusion from student surveys alone, it cannot reflect various other structural factors that affect parents and schools. Future studies should consider such aspects and other factors along with student surveys. Lastly, future studies should improve the accuracy of the model through data mining algorithms such as bagging, boosting, and random forest through ensemble methods to complement the disadvantages presented by decision trees.

5 References

1. Berrington, A., Roberts, S., & Tammes, P. Educational aspirations among UK young teenagers: Exploring the role of gender, class and ethnicity. British Educational Research Journal, 42(5), 729–755 (2016)
2. Berzin, S.C. Educational aspirations among low-income youths: Examining multiple conceptual models. Children & Schools, 32(2), 112–124 (2010)
3. Cupples, L.A., Bailey, J.N., Cartier, K.C., Falk, C.T., Liu, K.Y., Ye, Y., Yu, R., Zhang, H., & Zhao, H. Data mining. Genetic Epidemiology, 29(1), 103–109 (2005)
4. Deosaran, R.A. Educational aspirations: Individual freedom or social injustice? Interchange, 8(3), 72–87 (1977)
5. Kass, G. An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29(2), 119–129 (1980)
6. Khattab, N. Students′ aspirations, expectations and school achievement: What really matters? British Educational Research Journal, 41(5), 731–748 (2015)
7. Milanović, M. & Stamenković, M. CHAID decision tree: Methodological frame and application. Economic Themes, 54(4), 563–586 (2016)
8. Murphy, K.P. (2012). Machine learning: A probabilistic perspective. Cambridge, MA: The MIT Press.
9. Witten, I.H., Frank, E., Hall, M.A., & Pal, C.J. (2016). Data mining: Practical machine learning tools and techniques. Burlington, Massachusetts: Morgan Kaufmann.

Analysis on the Effects of Facility General Efficiency on Productivity through Non-operating Time Analysis of Small and Medium Manufacturing Processes[1]

Jae Sung Kim[1], Jeong Hyun Lee[1], Jinmyeong Cho[1], Wan Sup Cho[1]

[1] Dept. of Big Data, Chungbuk National University, 1, Chungdae-ro, Seowon-gu,
Cheongju-si Chungcheongbuk-do, Korea
{comkjsb, tosca2000cdx, britzegs, wscho63}@gmail.com

Abstract. In order to improve productivity in manufacturing enterprises, it is very important to secure the stability and availability of production facilities. To do so, systematic maintenance of facilities is highly necessary. Manufacturing facilities are closely related to productivity and profitability, and if unexpected downturns occur in manufacturing facilities, they will have significant impacts and losses in economic terms such as time / cost. Therefore, in order to improve the productivity of the manufacturing enterprise, it is necessary to analyze the factors hindering the productivity and systematically manage the facility efficiency. Therefore, in this paper, we analyzed the effect of facility operation data on the productivity from the point of view of overall equipment effectiveness through the analysis of facility operation data such as facility maintenance history, history of repairs, non - operation status data, and production information data.

Keywords: Productivity improvement, non-operation time analysis, overall equipment effectiveness (OEE), total productive maintenance (TPM), facility efficiency improvement plan

1 Introduction

It is very important to have the stability and availability of production facilities because unexpected non-operation situation of manufacturing facilities is very influential and loss in terms of economy such as time / cost. To do this, it is essential to eliminate productivity bottlenecks through unplanned non-uptime analysis of the plant, and to ensure productivity by increasing the efficiency of the production line facilities through optimization of the plant operation conditions.

The OEE of TPM can be decomposed into three types of facility productivity scales: time operation rate, performance operation rate, and yield rate depending on the type of loss such as facility stoppage time, equipment speed degradation, and product failure

[1] This work was supported by the Technology Innovation Program(or Industrial Strategic Technology Development Program)(20001798, Development of Integrated Management System for Intelligent Shoes Factory) funded by the Ministry of Trade, Industry&Energy (MOTIE, Korea)

due to facilities. OEE is a representative strategy in companies that can achieve both quality and cost by eliminating time loss, speed loss, and bad loss through these measures. [1]

Losses on the production site can be classified as non-loading loss, non-working loss, and non-operating loss in terms of production efficiency. Analysis and countermeasures are needed to improve production efficiency.[2]

Therefore, this paper analyzes the non-operation time using the field data of the automobile parts manufacturing process based on the total efficiency industry standard, which is one of the measurement indexes of TPM (Total Productive Maintenance) system, I would like to suggest ways to identify problems and improve productivity.

2 Precedent research

2.1 Overall Equipment Effectiveness

Overall Equipment Effectiveness (OEE) is a performance indicator that collects the status of a facility and verifies how efficiently the facility is being utilized. It measures the availability, efficiency, and quality of facilities. [3,4,5,6]

It is calculated by taking into consideration the effect of the six major losses such as equipment failure, item replacement, setting and adjustment loss, idling and speed drop, process failure, and yield reduction (Figure 1), which significantly degrade facility overall efficiency (OEE). By analyzing and improving the factors where the losses occur, it is possible to increase the efficiency of the equipment and improve the productivity.

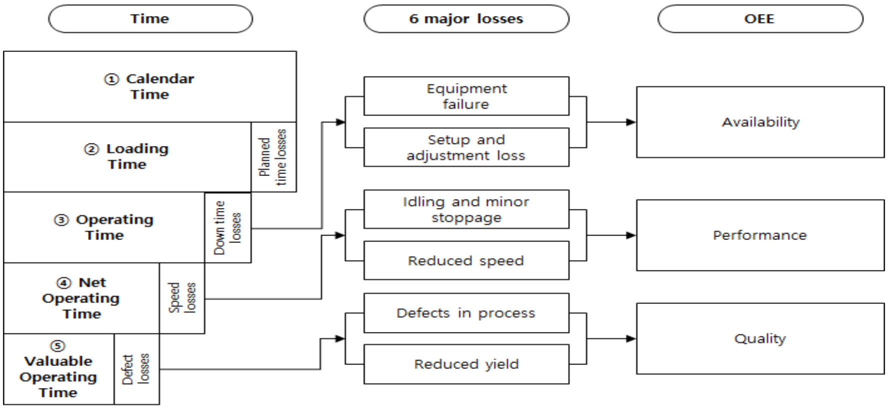


Fig. 1. OEE based on six major production losses

If losses in terms of facility operation level is classified, non-operating loss, no-loading loss, nonworking loss, and loss due to failure can be classified as shown in Table 1.

Type	Content
Calendar Time	It is expressed as the total time that the facility can be utilized based on 24 hours a day, 365 days a year.
Loading Time	It is the time in calendar time excluding planned scheduled time losses due to holidays, rest, training, scheduled maintenance, and so on
Operating Time	It is the time used for production of actual products, except Downtime Losses, due to equipment failure, stoppage due to equipment failure, setup, adjustment, change of breed
Net Operating Time	It is the time when machining is performed at a constant speed with respect to the operation time. It is the time excluding the time of loss (Speed Losses) due to deterioration of facility performance such as instant stop, idling, speed decrease.
Valuable Operating Time	It is the time used to produce superior products, except for the rework time due to defects.

Table. 1. Type of Production losses

2.2 Loss Structures to reduce OEE

There are 6 Losses (Table 2) that hinders the total efficiency of the facility during the operation of the plant. This includes Breakdown, Set-up and adjustments, idling, minor stoppages, Quality Loss, Reduced yield.

Type	Content
DOWNTIME	- Breakdown - Set-up and adjustments
SPEED	- Idling, minor stoppage - Reduced speed
QUALITY	- Quality losses - Reduced yield

Table. 2. Six Big Losses model proposed by Nakajima[7]

By utilizing this Loss structure, it is possible to analyze the Loss configuration such as Loss amount, Loss time, and Loss share, and to derive a Loss to be improved.

2.3 Facility productivity index

1) load factor (Benefit)

To demonstrate the productivity of a facility, two indicators of facility overall efficiency (OEE) and load factor should be considered together. The load factor can be calculated from the following formula.

* Benefit (formula 1)
= (Calendar Time - Planed time losses) / Calendar Time x 100

2) Overall Equipment Effectiveness (OEE)

The Overall Equipment Effectiveness (OEE) is calculated as the product of three factors: Availability, Performance, and Quality. Tracking the factors calculated by the Overall Equipment effectiveness can tell if the plant has more downtime than expected, the plant is operating at a slower rate, has a slight pause, or has created more defects [9]. Availability, also known as time of operation, indicates how much the machine is used for a valuable function against the plan. It is used as an indicator of how much time is wasted due to breakdown, preparation, replacement, etc. for the work time for a planned work instruction (formula 2). Performance is also called performance utilization rate. The efficiency is calculated by subtracting the theoretical time required for the input, and it includes idle, momentary stop, slowdown, etc(formula 3). Quality is calculated as the ratio of the quantity of good products to the actual quantity of production, excluding nonconforming products (formula 4).

* Availability (formula 2)
= (Loading time − Downtime) / Loading time x 100
* Performance (formula 3)
= (theoretical cycle time x Processed amount) / Operating time x 100
* Quality (formula 4)
= (Processed amount − Defect amount) / Processed amount x 100
* OEE (formula 5)
= Availability x Performance x Quality

The Overall Equipment Effectiveness (OEE) is calculated as shown in formula 5. As shown in Table 3, the World Class level for OEE is 85%, availability 90%, performance 95.0% quality 99.0%

OEE Factor	World Class
AVAILABILITY	90.0%
PERFORMANCE	90.0%
QUALITY	99.9%
Overall OEE	85.0%

Table. 3. Defining World-Class [9]

The productivity of the facility is calculated as "Benefit (load ratio) * OEE (Overall Equipment Effectiveness)". Improvement of load factor requires expansion of business and shortening of plan stopping time, and OEE is necessary to establish efficiency improvement measures in production technology, production and maintenance.

3 Experiment contents and model design

3.1 subject of experiment

The subject of experiment is Damper Pulley. Damper Pulley is a pulley mounted on an engine crank-shaft. It transmits rotational power to a water pump, alternator, A / C compressor through a belt and absorbs rotational vibration of crank-shift to prevent shaft breakage.

3.2 Data Set

We built a collection database for the six-month preservation period, non-operation history, and production daily data from July 2018 to December 2018. In addition, the analysis of the Ross structure and the improvement points of where the loss occurred through the process center non-operation time analysis were derived, and the effect of the Overall Equipment Effectiveness (OEE) on the productivity was analyzed.

In the maintenance job daily report, work start time, work end time, workplace, equipment, failure phenomenon, measures, worker, plan date, production plan number, part number, product name, production plan quantity, work instruction quantity, Quantity, and defective quantity. Non-operation details are divided into management, facility, item, material, quality, and others according to the Non-operation cause for each operation line, and the downtime is managed every minute.

The workshop operates 24 hours a day for 3 shifts per person and operates 5 days a week, so the total time that the 24-hour standard equipment can be utilized excluding the holiday, that is, the operating time is set as Calendar Time. The cycle time is set to calculate the plant overall efficiency (OEE) for the process line by adding the cycle time of the bottleneck facility during the process. Based on the fault phenomenon and the measures taken in the facility maintenance daily report, it is divided into the target facilities and the faulty parts, and Downtime Type is classified as equipment failure, item replacement, plan maintenance, and experiment development.

4 Analysis and Results

In order to analyze the correlation with the quantity produced in the process, we analyzed the correlation between the load factor and the production quantity, the planned downtime and the production quantity. We also derived availability, efficiency, and quality indicators to derive OEE. This study analyzed the effect of OEE on production.

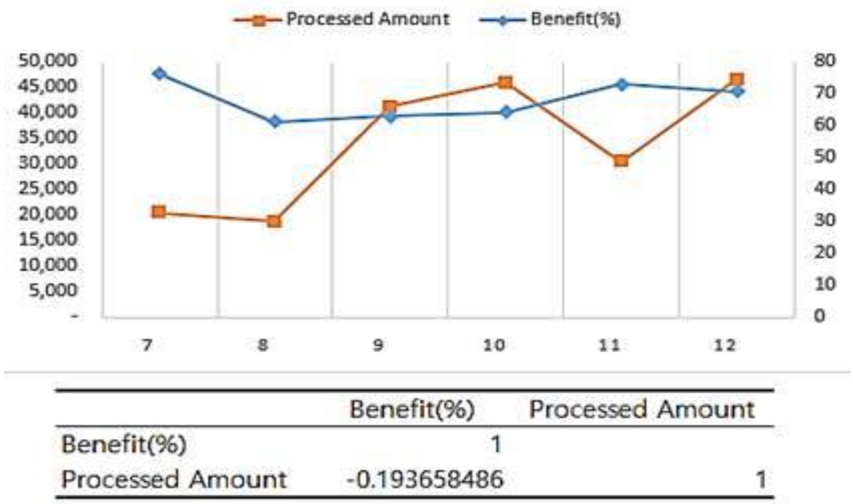
4.1 Correlation Analysis between the Benefit and the Processed Amount

The average Loading Time and the average Loading Rate for the 6 months excluding the planned time losses were 29,850 minutes and 68.01%, respectively. The load factor is derived from Formula (1).

Month	Calendar Time(min)	Planned time losses(min)	Loading Time(min)	Benefit (%)	Processed amount(ea)
7	43,200	10,290	32,910	76.18	20,497
8	44,640	17,310	27,330	61.22	18,766
9	43,200	15,990	27,210	62.99	41,122
10	44,640	16,050	28,590	64.05	45,941
11	43,200	11,700	31,500	72.92	30,483
12	44,640	13,080	31,560	70.70	46,541
	Average		29.850	68.01	33,892

Table. 4. Loading Time and Benefit

As a result of correlation analysis between monthly load factor and production yield, correlation coefficient value showed a low correlation with -0.193. These results in Figure 2 demonstrate that the productivity is not explained by various loss factors even though the load factors do not vary greatly. Therefore, additional Loss factor needs to be discovered through analysis of Loss structure.



	Benefit(%)	Processed Amount
Benefit(%)	1	
Processed Amount	-0.193658486	1

Fig. 2. Benefit vs Processed amount

4.2 Correlations between Planned time losses and Production Quantities

Planned time losses excluding holidays are listed in order of inventory, quality awards, and meetings. Health screening, roundtable discussion and sexual harassment prevention training at work Table 5.

Contents	Time (min)	Ratio (%)
listed in Inventory	600	66.67
quality awards	120	13.33
meetings	90	10.00
Health screening	30	3.33
roundtable discussion	30	3.33
Sexual harassment prevention training at work	30	3.33

Table. 5. Planned time losses

As a result of the analysis of the monthly unit downtime analysis, the planned downtime was in the order of July, October, November, September, and August, and the correlation analysis between the downtime and the production yield showed a low correlation (Table 6). This implies that the downtime does not significantly affect productivity.

	Planned Time Loss	Processed Amount
Planned Time Loss	1	
Processed Amount	0.293710453	1

Table. 6. Planned time losses vs Processed amount Correlation Analysis

4.3 OEE Analysis through Quality Indicators

Since the Overall Equipment Effectiveness (OEE) is calculated as the product of the three factors of availability, performance, and quality, the availability index, the efficiency index and the quality index should be calculated to obtain the OEE.

4.3.1 Availability

Time operation rates can be obtained through a measure of availability. The operating time (except for Downtime Losses) at loading time is shown in Table 7 and the average operation time is 29,297 minutes.

The availability for time availability is derived through Formula (3). As a result of the analysis, the average availability for 6 months was 98.13% as shown in Table 6. In addition, correlation analysis result between monthly availability and production yield showed a low correlation with -0.01. It is analyzed that the actual operation time excluding the maintenance loss (Downtime Losses) does not have a significant effect on productivity. Therefore, it is necessary to identify core loss factors through Loss structure analysis.

World Class : 90%

Month	Loading Time(min)	Downtime losses(min)	Operating Time(min)	Availability (%)	Processed amount(ea)
7	32,910	650	32,260	98.02	20,497
8	27,330	590	26,740	97.84	18,766
9	27,210	650	26,560	97.61	41,122
10	28,590	540	28,050	98.11	45,941
11	31,500	320	31,180	98.98	30,483
12	31,560	570	30,990	98.19	46,541
Average	29,850	533	29,297	98.13	33,892

	Processed Amount	Availability (%)
Processed Amount	1	
Availability (%)	-0.010427667	1

Table. 7. Availability vs Processed amount Correlation Analysis

As Table 8 shows, downtime loss is due to equipment failure, during downtime due to replacement, planned maintenance and development Equipment failure is 1,630 minutes (52.41%), which is the largest cause of downtime.

Contents	Time (min)	Ratio (%)	Number of downtimes	Ratio (%)
Equipment failure	1,630	52.41	27	62.79
Replacement	1,050	33.76	12	27.91
Planned maintenance	320	10.26	3	6.98
Development	110	3.54	1	2.33

Table. 8. Downtime losses

However, as shown in Table 9, the correlation coefficient between equipment failure time and production shows a low correlation of 0.275. You can see that equipment failure time is not large enough to affect productivity.

Month	Equipment Failure Time	Processed amount(ea)
7	230	20,497
8	330	18,766
9	320	41,122
10	270	45,941
11	140	30,483
12	340	46,541

	Equipment Failure	Processed Amount
Equipment Failure	1	
Processed Amount	0.275390038	1

Table. 9. Equipment Failure vs Processed amount Correlation Analysis

4.3.2 Performance

Performance is also known as performance utilization and measures yield during actual uptime. Downtime is excluded during actual uptime. Performance utilization and availability are as follows: (formula 4). Theoretical Cycle Time is During the operation of several lines, the theoretical cycle time of the bottleneck facility. The equipment and theoretical cycle times are calculated as follows. As a result of the analysis, the average efficiency performance over the six months is as follows. As shown in Table 10, the monthly deviation is 70.03%, which is insufficient compared to 95% of the global level.

Month	Operating Time(min)	Processed amount(ea)	Performance (%)
7	32,260	20,497	38.12
8	26,740	18,766	42.11
9	26.560	41,122	92.90
10	28,050	45,941	98.12
11	31,180	30,483	58.66
12	30,990	46.541	90.11
Average	29,297	33,892	70.03

Table. 10. Performance

According to interviews in the workplace, if the alarm is frequent and suddenly stops, the efficiency of the facility will depend on the skill of the driver. In the case of an alarm, a momentary stop is often ignored because it can be restored to its original state. Extension of this problem. It has been identified as an important key cause of productivity problems.

	Processed amount(ea)	Performance (%)
Processed amount(ea)	1	
Performance (%)	0.978429944	1

Table. 11. Performance vs Processed amount Correlation Analysis

The correlation coefficient between monthly performance utilization rate and production yield showed a high correlation coefficient of 0.978 (Table 11). This shows that there is a close correlation between the loss due to the instantaneous stoppage, ball rotation, and speed drop due to the alarm generation. It is analyzed that systematic management is needed.

4.3.3 Quality

Quality is calculated as the ratio of the number of good products to the actual quantity of production, excluding nonconforming products. The quality index is derived from Formula (4). As a result of the analysis, the average quality for 6 months is 99.99% as shown in Table 11, and it is analyzed that it is well managed.

Month	Processed amount(ea)	Defect amount(ea)	Good amount(ea)	Quality(%)
7	20,497	1	20,496	100.00
8	18,766	6	18,760	99.97
9	41,122	22	41,100	99.95
10	45,941	6	45,935	99.99
11	30,483	104	30,379	99.66
12	46.541	3	46,538	99.99
Average	33,892	23.67	33,868	99.99

Table. 12. Quality

4.3.4 Overall Equipment Effectiveness Analysis

The OEE is derived from Formula (5). As shown in Table 13, the average total equipment efficiency (OEE) of the workplaces was 68.71%.

Month	Availability (%)	Performance (%)	Quality (%)	OEE (%)
7	98.02	38.12	100.00	37.37
8	97.84	42.11	99.97	41.19
9	97.61	92.90	99.95	90.63
10	98.11	98.12	99.99	96.26
11	98.98	58.66	99.66	57.58
12	98.19	90.11	99.99	88.47
Average	98.13	70.03	99.99	68.71

Table. 13. OEE

Analysis of the average total efficiency of the facilities shows that although the availability and quality indicators are good, the performance index is low, so the overall average efficiency of the facilities is hindered.

4.3.5 Impact of OEE on production

Table 14 summarizes the relationship between OEE, loading time, operating time, and output (quantity of good goods).

Month	Loading Time(min)	Operating Time(min)	OEE(%)	Good amount(ea)
7	32,910	32,260	37.37	20,496
8	27,330	26.740	41.19	18,760
9	27,210	26.560	90.63	41,100
10	28,590	28,050	96.26	45,935
11	31,500	31,180	57.58	30,379
12	31,560	30,990	88.47	46,538
Average	98.13	29,297	68.71	33,868

Table. 14. OEE vs Good amount

Correlation analysis of the facility's overall efficiency (OEE) and output (good product quantity) resulted in 0.979 as shown in Table 15.

	OEE(%)	Loading Time(min)	Operating Time(min)	Good amount(ea)
OEE(%)	1			
Loading Time(min)	-0.064	1		
Operating Time(min)	0.020	-0.057	1	
Good amount(ea)	0.979	0.027	0.187	1

Table. 15. OEE vs Good amount Correlation Analysis

This shows that the increase in OEE leads to the increase in the production. In a study by Hyojoon Jang [10], there is a research result that production efficiency increases by 6.2% and operating profit by 95.5% when the OEE increases from 81.6% to 82.5%. Therefore, it is necessary to identify the factors that impede OEE and make efforts to improve them.

5 Conclusion

Changes in OEE have a direct impact on productivity. Therefore, productivity and quality can be improved by defining and analyzing loss and efficiency impeding factors in terms of facility and quality. Overall equipment efficiency is determined by availability, performance and quality, and improves plant efficiency to maximize plant capacity. that's You can improve your productivity by analyzing and eliminating the loss factors of production site equipment such as downtime loss, speed loss and quality loss.

The results of the analysis show that equipment failure time has little effect on productivity. It can be seen that the equipment failure time is not large enough to affect productivity. The key factor that affects the actual productivity is that the cause of small trouble such as instant stop which does not appear in the work day greatly hinders productivity efficiency. In order to reduce idling, momentary loss, and speed loss, it is necessary to maintain independence prevention and to educate and skillful workers. In addition, it should be able to take immediate action and manage it in case of alarm or instant stop.

We will actively collect facility data, analyze hidden patterns and analyze associativity, and proactively respond to key issues that hamper productivity, thereby ensuring facility efficiency and contributing to product productivity and quality improvement.

References

1. Choi, Sungwoon.: Development and Analysis of Fuzzy Overall Equipment Effectiveness (OEE) in TPM, Journal of the Korea Management Engineers Society Vol. 23 No. 4,pp. 87 – 103(2018)
2. Jae Kung Lee, Seung Woo Lee,: Downtime tracking for small-medium sized manufacturing company using shop floor monitoring, Journal of the Korea Industrial Information Systems Society (79):65-72(2014)
3 . I.P.S. Ahuja and J.S. Khamba.: Total productive maintenance: literature review and directions, International Journal of Quality & Reliability Management (2008)
4. Punjabi University, Patiala, India. International Journal of Quality & Reliability Management Vol. 25 No. 7, pp. 709-756(2008)
5. Dong Su Kim, Duk Hee Moon :A Case Study of Comparing the Measuring Methods for Workloads of Resources in a Manufacturing Processes of Semiconductor-Parts, Korea Society for Simulation 49-58(2011);
6. S.K. Cha,J.Y. Kim,J. Y. Yoon.:Agent system for implementation of the standard based OEE, KoreanSociety for Precision Engineering 1475-1476(2013)
7. Nakajima, S.: Introduction to TPM: Total Productive Maintenance. Productivity Press;(1988).
8. Jong Yup Beak, Yoon Jin Kang, Kung Sik Kang.:Improving Overall Equipment Effectiveness(OEE) in Korean Small and Medium Manufacturing Industries, Korea Safety Management & Science, p. 219-230(2010)
9. http://www.oee.com/world-class-oee.html
10. Hyo Joon Hahm.: An Analysis of the Impact of Overall Equipment Efficiency on the Firm's Earnings,. The Korean Institute of Plant Engineering, Volume: 4 Issue: 2 p. 71-81. ISSN: 1598-2475((1999)

Impact Analysis of Contract Method on Awarding Price Ratio in Public Bidding

Tae-Hong Choi [1], Eun-Seon Choi[2], Phyoung-Jung Kim[3], Wan-Sup Cho[4]

[1] [4]Management Information Department, Chungbuk National University,
Cheongju, South Korea
[2] Big Data Cooperative Course, Chungbuk National University,
Cheongju, South Korea
[3] Computer Information Course, Chungbuk Provincial University,
Cheongju, South Korea
{ thchoi2 }@naver.com, { tmxk147 }@gmail.com, { pjkim }@cpu.ac.kr, { wscho }@chungbuk.ac.kr

Abstract. The goal of this study is to analyze the impact of the contract method for the awarding price ratio in the present public contract. In particular, this work collects and analyzes a lot of data on bid and contract for the purchase of goods, service and construction work by ordering organizations and companies in KONEPS. As an analytical model, the contract method is used as an independent variable. By using text mining, we collect big data and by using multi-dimensional method, we analyze almost 4.7 million big data on bid and contract in KONEPS for the last 20 years. The results of the analysis are as follows, (1) the contract method affects the awarding price ratio in both purchase of goods, service, and construction work. (2) Private contract was the highest, followed in the order of nominated competition, formal competition, and limited competition. (3) In the bidding type information was the highest construction work, followed in the order of service and purchase of goods. This research suggests that the corroborated analysis provides the efficiency in political perform.

Keywords: KONEPS, public bid, contract method, bid type, awarding price ratio

[1] Management Information Department, Doctor’s Program, Chungbuk National University/ PPS

[2] Big Data Cooperative Course, Master's Program, Chungbuk National University

[3] Chungbuk Provincial University, Professor.

[4] Department of Management Information System, Chungbuk National University, Professor, correspondent author.

1 Introduction

In the ever-changing public procurement market, procurement to support private enterprise activities has a great impact on corporate management and profit structure. If the public contract system and its procedures fail to function in a way that does not conform to the original purpose of the contract, resulting in a waste of the budget and adversely affecting the national economy(Moon, 2015). Although there are many researches on the awarding price ratio [5] of auction, which is one of the sources of revenue, there are relatively few researches on the factors influencing the awarding price ratio of the public contract, which is one of the annual expenditure sources. Therefore, it is necessary to conduct empirical analysis to identify the factors influencing the awarding price ratio using contract data to improve the contract system. The purpose of this study is (1) to examine the optimal price level by analyzing the structure of the awarding price ratio, which is not the manufacturing cost or the supply cost. (2) to help public works projects to be ordered with high-quality and profitability. This will be an opportunity to recognize the importance of the factors influencing the awarding price ratio for all public institutions that need to pursue financial efficiency at the same time. Comparative analysis of the awarding price ratio for each business, this work provides a comparable result to the existing results in the previous literature..

2 Related Research

2.1 Status of Public Contract in KONEPS

Most of today's transactions are being handled by electronic systems. In Korea, the PPS started the procurement EDI in 1997, and in 2002, it launched the electronic procurement service by establishing KONEPS[6](www.g2b.go.kr), web site that can search for bid information. As a result of the establishment and operation of a single window for nationwide procurement through the electronicization of the procurement process, resulting in over 70% of public transactions(PPS, 2019).

A public contract is different from a judicial contract in which a public institution is established for the public interest in pursuit of the public interest, and the public interest is established by the sign of the sign between the signatures. As a result, separate contractual laws[7] and regulations[8] have been enacted and operated. Public contracts are conducted in the order of the judicial contract, such as contract method determination→bidding announcement→ bidding→ awarder decision→ contract

[5] The awarding price ratio means 'awarding price / estimation cost'.

[6] KONEPS (Korea ON-line E-Procurement System) is a brand created by the Public Procurement Service (KPS) for the international community in order to secure the uniqueness of Nara Changter, a national e-procurement system, and to communicate its excellence to the whole world(PPS, press release, 2007.3.12.).

[7] State Contract Law, 2018 Revision Implementation, Local Contract Law, 2018 Revision Implementation,

[8] The Ministry of Strategy and Finance, revised in 2018, bidding and contract execution standards of local governments, criteria for awarding bids for local government bidding, Ministry of Public Administration and Security, revised in 2018.

signing→ contract fulfillment→ payment of payment, but the kind is classified according to contract object, contract type, competition methods, and bidding types.

2.2 Studies on the Factors Affecting the Awarding Price Ratio

A result of analyzing the bidding data of public IT procurement (Eg, Linear Regression, Support Vector Regression, and Random Forest, which are most commonly used in practice, the conclusion of the contract(contract method) was confirmed as a factor influencing the awarding price ratio(Kim, 2017). In the case of the DT model, the limited competition contract and the formal competitive contract have a high positive correlation with the awarding price ratio, but the private contract has negative correlation (Kim, 2019). In the case of the limited competition, the contract method may affect the number of bidders and the awarding price ratio, as in the case of the study that companies located in other cities and provinces are excluded from bidding and the competition level is lowered in preparation for formal competition(Cho et al., 2014; Choi et al., 2013; Kim et al., 2011). In the case that the awarding price ratio is lowered through self-efforts such as promoting the private contract business as an open competition contract, collecting the business of various departments, ordering a bundle, or applying the lowest price bidding system(Bae et al., 2013). In addition, the awarding difference between the awarder and the contractor's own efforts such as changing the contract method at the execution stage of the contract is recognized as the budget-saving result(Korea Construction Association, 2018; Cho et al., 2014).

3 Analytical Model and Result

3.1 Analytical Model and Operational Definition

Considering the influence factors used in the previous study in Section 2.2 and the characteristics of the KONEPS public contract, which is the subject of this study, the research model for the analysis of the awarding price ratio assumes that the Contract Method is an independent variable <Figure 1> shows the study model with the winning price ratio as a dependent variable.

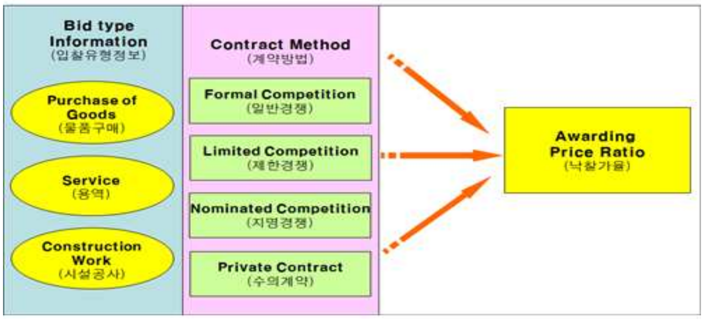


Fig. 1. Analytical Model.

<Table 1> summarizes the operational definitions and measurement items of the variables required in this study.

Table 1. Operational Definition of Variables and Metrics.

Division	Variables	Operational Definition	Metrics
Independent Variables	Contract Method	How to sign a contract	Types of contract method adoption
Dependent Variable	Awarding Price Ratio	Ratio of awarding price to estimation cost	The amount of awarding price relative to estimation cost

3.2 Collecting Data and Test method

This study mainly collected data by using open API of public data portal (www.data.go.kr/). The collected data is analyzed using Big Data and Multidimensional Analysis. In other words, at the top of <Figure 2>, the data warehouse (DW) is extracted from KONEPS, and the OLAP Cubes are constructed and analyzed according to the analysis purpose. Data collected using the open API of public data portal (www.data.go.kr/) was stored in DW after proceeding with ETL using MySQL 8.0 ver. We extracted the data from the database using the query, and separated the data extracted by using R.

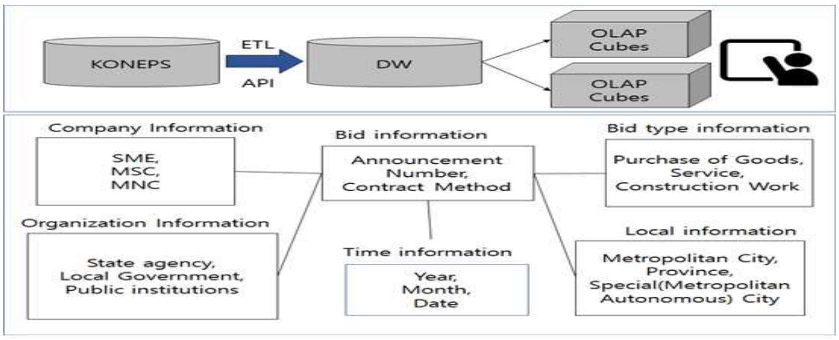

KONEPS
ETL
API
DW
OLAP Cubes
OLAP Cubes
Company Information
SME, MSC, MNC
Bid information
Announcement Number, Contract Method
Bid type information
Purchase of Goods, Service, Construction Work
Organization Information
State agency, Local Government, Public institutions
Time information
Year, Month, Date
Local information
Metropolitan City, Province, Special(Metropolitan Autonomous) City

Fig. 2. Data warehouse building process and star schema structure.

3.3 Results

In this section, we analyze the influence of contract method and awarding price ratio in the purchase of goods, service, and construction work. <Table 2>, <Figure 3>, and <Figure 4> show the total number of bidding and the mean awarding price ratio according to the contract method for the purchase of goods. The contract method is classified into formal competition, limited competition, nominated competition, and private contract based on the contract method of the purchase of goods ordered using KONEPS. There were many missing values (NA) of the contract method in collected data, and there were few cases mixed with awarding method. Outliers that are difficult to classify contract methods or away from other observations have been removed. In the below figure, box diagram showing the distribution of the data, a small circle(°) indicating the extreme value is displayed. This data includes Outliers.

Table 2. Contract Method & Awarding Price Ratio(purchase of goods).

Contract method	Number	%	Awarding Price Ratio(%)			
			Mean	Min	Max	Med
Formal Competition	74,175	6.26	88.63	1.62	100	88.00
Limited Competition	404,299	34.11	87.35	1.08	100	87.76
Nominated Competition	13,010	1.10	89.13	9.76	100	88.53
Private Contract	692,479	58.43	90.40	1.57	100	88.11
the Others	1,236	0.10	79.22	50.24	99.28	81.02
Total	1,185,199	100				

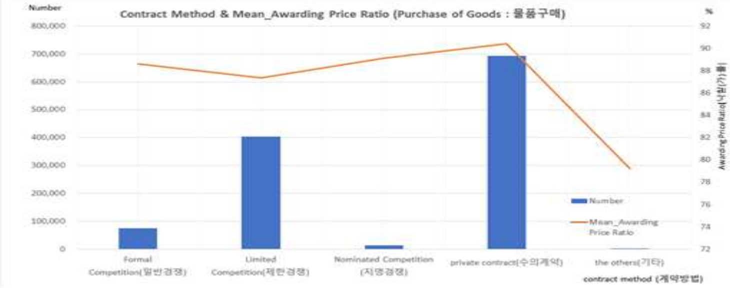


Fig.3 Budget & Mean_Awarding Price Ratio(purchase of goods).

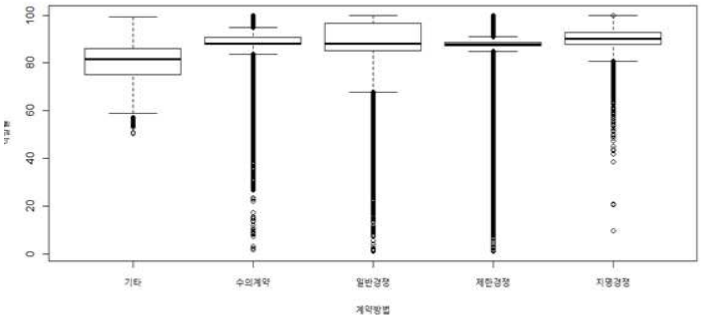


Fig.4. Contract Method & Mean_Awarding Price Ratio(Purchase of Goods)

<Table 3>, <Figure 5> and <Figure 6> correspond to cases of service contract. The following description is the same as the description of the purchase of goods and is omitted. In the above figure, the mean awarding price ratio is high in private contract and nominated competition. You can see that the pattern is similar to the purchase of goods.

Table 3. Contract Method & Awarding Price Ratio(Service)

Contract method	Number	%	Awarding Price Ratio(%)			
			Mean	Min	Max	Med
Formal Competition	87,913	6.57	88.49	1.01	100	88.05
Limited Competition	453,904	33.93	87.47	1.08	100	87.76
Nominated Competition	21,567	1.61	90.20	9.76	100	90.24
Private Contract	773,012	57.79	90.38	1.57	100	88.17
the Others	1,332	0.10	79.53	50.24	99.28	81.59
Total	1,337,728	100				

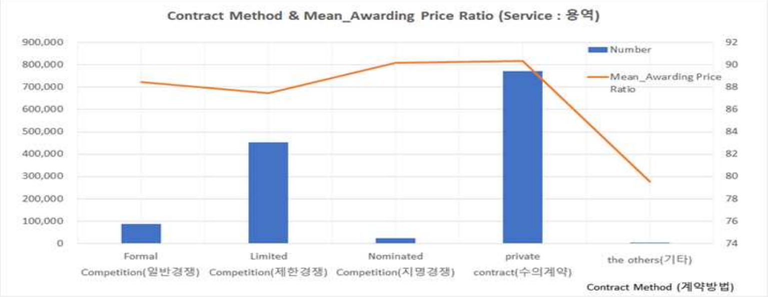


Fig.5. Contract Method & Mean_Awarding Price Ratio(Service)

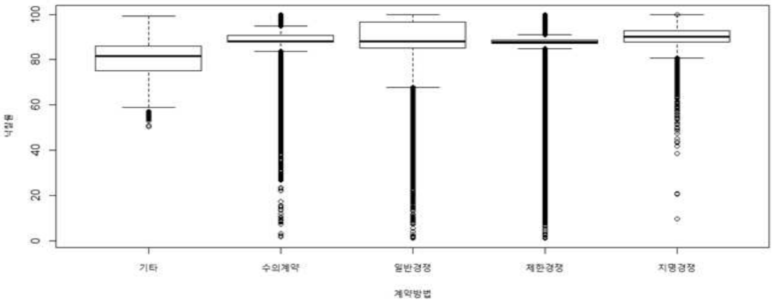


Fig.6. Contract Method & Awarding Price Ratio scatterplot(Service)

<Table 4>, <Figure 7> and <Figure 8> correspond to cases of construction work. The following description is the same as the description applied to the service and is omitted.

Table 4. Contract Method & Awarding Price Ratio(Construction Work)

Contract method	Number	%	Awarding Price Ratio(%)			
			Mean	Min	Max	Med
Formal Competition	87,331	6.54	88.49	1.01	100	88.05
Limited Competition	452,877	33.90	87.47	1.08	100	87.76
Nominated Competition	21,562	1.61	90.20	9.76	100	90.24
Private Contract	772,750	57.85	90.38	1.57	100	88.17
the Others	1,332	0.10	79.53	50.24	99.28	81.59
Total	1,335,852	100.				

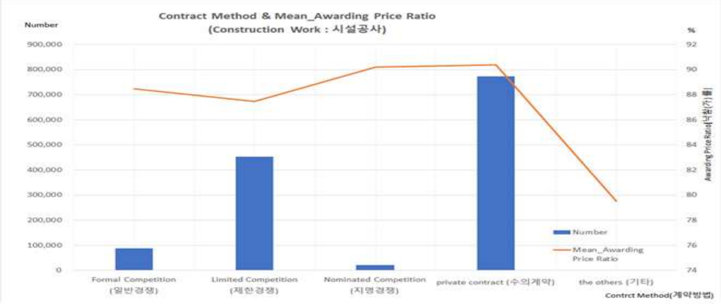


Fig.7. Contract Method & Mean_Awarding Price Ratio(Construction Work)

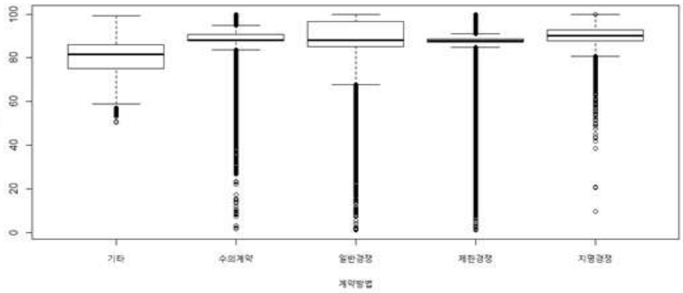


Fig.8. Contract Method & Awarding Price Ratio scatterplot(Construction Work)

<Table 4>, <Figure 7> and <Figure 8> correspond to cases of construction work contract. The following description is the same as the description of the purchase of the article and is omitted. As shown in <Table 2>·<Table 3>·<Table 4> and <Figure 3>·<Figure 5>·<Figure 7>, it can be seen that the contract method determines the awarding price ratio. The awarding price ratio was the highest in private contract, followed by nominated competition, formal competition, and limited competition. The awarding price ratio was the highest in private contract, followed by nominated competition, formal competition, and limited competition. It can be seen that the ultra-low price award of 1% in the awarding price ratio exists in the formal competition, the limited competition, and the private contract, and it is confirmed that there are many award over the latter half of the 80% by position of the mean value and median. In contrast to the previous research, it can be seen that the awarding price ratio of the limited competition against formal competition is low.

In the previous <Figure 4>·<Figure 6>·<Figure 8>, it can be seen that the number of outlier is significant in the order of limited competition, private contract, formal competition, and nominated competition, and the awarding price ratio is higher in private contract and nominated competition. In the bidding type information was the highest construction work, followed in the order of service and purchase of goods.

4 Conclusion

The results of this study are summarized as follows. The contracting method has an effect on the decision of the awarding price ratio. Among the contract methods, the awarding price ratio of the private contract was the highest, followed by the awarding price ratio in the order of nominated competition, formal competition, and limited competition. Next, policy implications are as follows. Institutional improvement is required in order to reduce private contracts and limited competitions and to promote formal competition.

The fact that formal competition, which is the principle of contract, is placed at an excessively low rate should be avoided. Compared with the purchase of goods with the lowest price and the service that did not, the higher awarding price ratio of the construction work is required to improve the contracting system.

As we have seen in the discussion above, this study is significant in the analysis that confirms that the contract method is the determinants of the awarding price ratio. However, the limitations of this study are as follows. A large amount of error data (such as missing values) has been found in the data collected at the KONEPS, and the process of correcting or eliminating them may affect the results of the data analysis, and further research is needed.

5 References

1. Bae, J. S. & B. S. Kim. The Evaluation and Improvement of Budget Saving Policy in Seoul Metropolitan Government (2013)
2. Cho, E. R. & J. S. Kim. Problems and alternatives of public ordering system. Issues & Diagnosis, - (150) : 1--25 (2014)
3. Choi, E. J., M. H. Kim, & Y. D. Kim. Analysis of public construction trends and implications. Construction issue focus, -(-) : 1--34 (2013)
4. Kim, C. Development of Bidding-Ratio Prediction Model for Public Information Technology Business Projects Using Data Mining Method. Yonsei University Graduate School of Engineering Master's Thesis (2017)
5. Kim, E. H. A Study on Determinants of Success Rate of IT Services Using Procurement Data from Public Procurement Service. Keimyung University Graduate School Master's Thesis (2019)
6. Kim, H. L., J. N. Rhee, & E. K. Lee. Determinants of the price of drugs purchased through open competitive bidding: Focusing on the case of one national hospital, health economics and policy research, 17(2): 1--23 (2011).
7. Korea Construction Association, Private Construction White Paper 2017. Seoul: Korea Construction Association (2018)
8. Moon, B. O. Empirical study on the estimation cost determination and successful bidder selection process in the government contracts. Daejeon University Graduate School Doctoral Thesis (2015)
9. PPS. Procurement Anniversary. Daejeon: Media Cheong (2019)

Design and Implementation of Book Sharing System with Open Source Software

Ahyun Cheon, Eun-Taek Lee, Jung-Hoon Kim
Department of Computer Science, Chungbuk National University, South Korea
cah0512@naver.com, dmsxodor2@naver.com, sjsrh96@naver.com

Abstract. Due to the development of information and communication technology, ownership of electronic devices is increasing. Also, Exposure of various media is increasing, and the reading volume of people is decreasing rapidly. Therefore, Book checks out and book sales rate are decreasing. Many studies show that reading affects learning efficiency, and not only promotes the accumulation of human capital, one of the essential factors driving economic development but also has an economic impact, such as helping nurture creative talent that creates knowledge. As a result, People should recognize this and realize the need for reading, and encourage reading to give them an incentive to improve reading habits. Therefore, we devised a book exchange project through the making of people's application form, in which information can be stored in the DB, and book gifts can be exchanged randomly. In addition, additional web services can generate interest in reading.

Keywords: reading, reading effect, web service

1 Introduction

According to a survey on the current reading of the public, one in 10 young people in their 20s and 30s do not read a single book throughout the year. Annual reading is also decreasing over the years. On the other hand, the consumption of video content is soaring. According to a survey conducted by the Ministry of Culture, Sports and Tourism, for adults, the annual reading dropped from 9.1 in 2015 to 8.3 in 2017, and elementary, middle and high school students also declined to 28.6 in 2017 from 29.8 in 2015. However, according to data from 'Wize App', which examines app analytics and user trends, YouTube usage time increased from 11.7 billion minutes in September 2016 to 20.6 billion minutes in September 2017 and 29.4 billion minutes on September 2018. In addition, modern people often replace reading reviews with video reviews without reading books, resulting in the creation of a book review, and this has led to more people replacing the book itself. But in this case, people don't have their own opinions or ideas about books, and only remember the opinion that the video producer gave them or the opposite opinion. This creates a habit that can degrade your thinking ability. The video cannot give you enough time for your curiosity to mature because you can quickly find the answer to the question. Therefore, one's concentration or thinking ability can be reduced by information pouring out without being able to concentrate on one thing. In a modern society where smartphone and Internet use is overused, the way to get information is very easy to get without thinking and doing anything. It is also unclear whether the information obtained is accurate.

Literacy is falling as the reading volume decreases, and the reliance on video increases. In today's society, students have a fairly good digital literacy, but their ability to read and interpret paper books is remarkably poor. Also, as the reliance on the Internet or smartphones increases, psychology says it increases depression, anxiety, and aggression, and its short-term memory ability decreases. Therefore, it is necessary to increase the volume of reading and to motivate people to appreciate reading. So we try to turn people's perception of reading into something more interesting and fun through interesting events on the website.

1. We will generate people's interest through book exchange projects and boost reading sales as well as reading volumes. Gifts one's book to a stranger, and also to himself a gift arrives, arousing curiosity about the book so that he can start reading quickly. Once you complete the application by entering your own information for receiving gifts on the application form, you can randomly receive contact information and address information from strangers. You can send him a book gift, and you can get a book exchange project done by getting a book from another person.
2. Through people's book evaluations and recommendations, it increases exposure to books, exchanges of views on books and creates opportunities to enhance thinking and judgment. It allows people to look at short, book-recommended articles and have an interest in popular magazines and books that people have read universally.
3. Through book donation activities, one can feel individual achievement with a positive attitude toward donation and create a healthy donation culture. You can donate a collection of fairy tales, a collection of great books, or books you've read and left behind as a child, where you need help. You can also have time to revisit books that you don't need.

The paper proceeds as follows. Section 2 discusses the related study. Section 3 explains the proposed method, along with data pre-processing and data analysis. Section 4 presents the implementation results. Section 5 concludes the paper and marks future work.

2 Related Study

In this section, we discuss the related work. To successfully deliver Web services, we need to understand the need for reading fundamentally. There are already many papers on the necessity and effectiveness of reading.

First, a study by researchers at Sussex University in England found that reading books works well when relieving stress. If you read a book for six minutes or so, your stress decreases by 68 percent, your heart rate decreases, and your muscles relax. It is not important to read any book in this situation. The act of reading a book itself helps relieve stress by escaping from the stress and falling into the writer's story.

Second, according to a comprehensive education study conducted by Eulji University in South Korea, reading volume directly affects the learning performance of university students. The results of this study are consistent with the research that shows a significant correlation between reading motivation and academic performance during the preceding study, and that the broader reading of ordinary times improves knowledge and thinking skills, and thus the confidence that comes from it increases learning performance. Therefore, reading can result in acquiring complex information, pursuing one's dreams, and leading one to live a high-quality life, opening up the mental world and leading one to live a broader life.

To increase the need for reading, various services are now available. ‘Fly book’, a data-based online book platform, is a service that helps you achieve your target reading volume by setting a reading target and registering the number of books you want to read during the year. In addition, when registering a book after reading, it shows the status of achievement in real-time graphs based on the target reading volume and provides monthly reading trend graphs, etc. And the monthly subscription-based reading app, ‘Millie's library’, is gaining popularity, paying for it with monthly installments so that people can read a wide range of books every month. The service, which aims to become friendly with reading, and even the platform introduced earlier, are all providing people with active reading activities.

However, the platform, called "Fly book", can be challenging to act on because it only sets reading targets and has to get books for itself. The biggest reality is that online and offline bookstores and libraries are always nearby, but they don't actually go there. In addition, the 'Millie's library' service can cause users to worry about using it because it requires users to pay a certain amount of money to use it. In our development system, users are not required to purchase books or pay for them. Just send the book you read as a gift to others. In other words, the system activates neglected books. If there are many books that have been read interestingly in the past but have not been re-opened after reading them, give them to others through the book exchange project. And you'll get a book gift from someone you don't know too. The book I received as a gift makes me want to read about it in curiosity. This provides an opportunity for reading to become closer. It also allows you to read people's book recommendations on Web services, get them interested in books, and feel like reading them on their own. Finally, you can use the book donation service on the Web to donate books where you need help. This is an activity that can satisfy an individual's self-efficiency and sense of accomplishment. s

3 The database and program structure

This section describes the relationship between the main core functions and the database using the MVC-based Model 2 architecture, which is the basic structure of Web services.

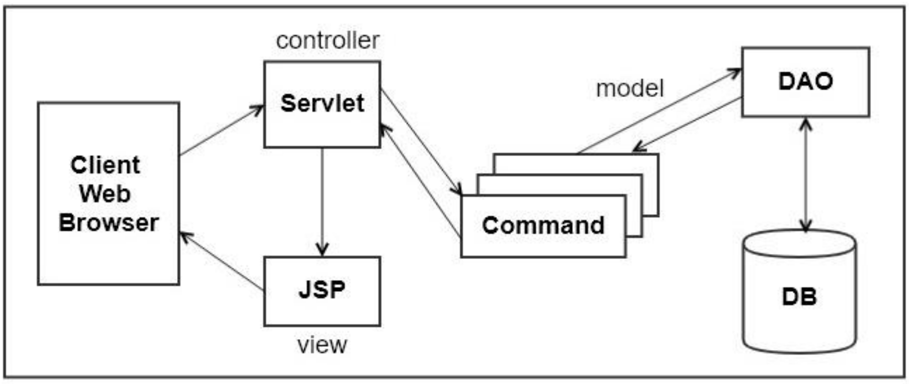


Fig. 1. MVC-based Model 2 Architecture

Our Web service that implements MVC patterns consists of three modules: Model, View, and Controller. Fig. 1 shows the web structure of MVC-based Model 2 Architecture. The model means Business Logic and is implemented using Java Beans and normal classes. Typical is DAO, DTO, and Command classes. View stands for Presentation Logic and uses JSP. Controller means a logic that properly manages Model and View and is implemented using servlet. The model handles business logic, which means actual operations related to client requests. The request is modularized using command patterns and implemented as a DAO class and DTO class for database interworking. The view is responsible for the presentation logic that responds to the client as a result of the model's execution and is implemented as JSP.

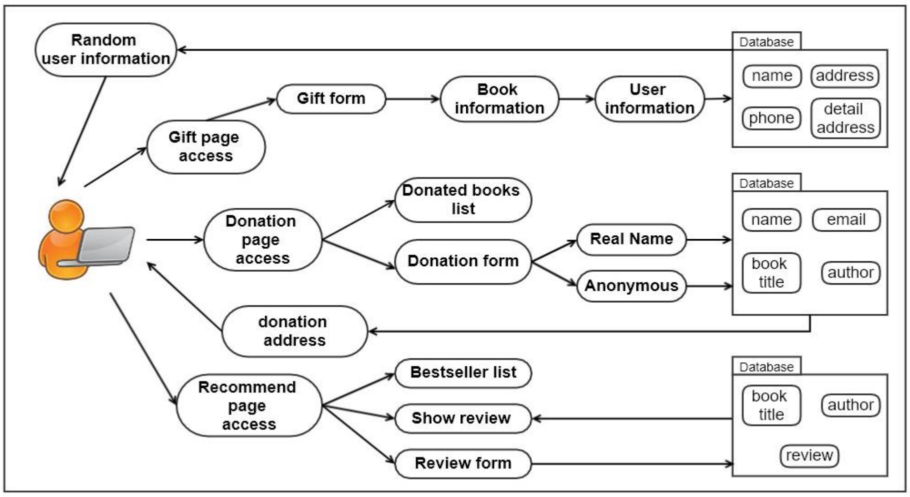


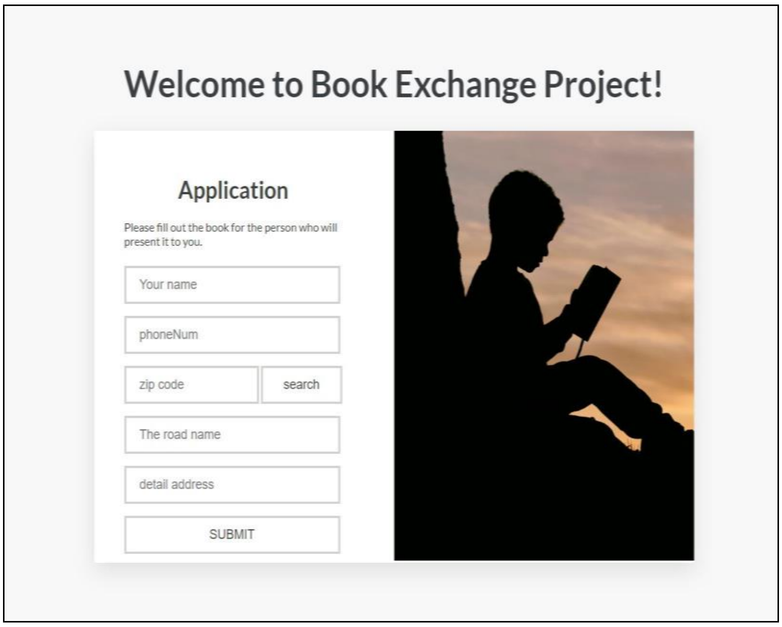
Fig. 2. Functional Architecture

Fig. 2 is an architecture that demonstrates the implementation of the representative functions of our project. The page is classified as a gift, donation, and recommendation. When the user accesses the gift page, there is a form that enters the applicant's information. After you enter the information for the book, you want to present, enter your name, address, and phone number, and the information is stored in the database. When the application is complete, the information of the person who will send the book appears on the screen.

At this time, one random person is selected from the information stored in the database, except his or her information. The information of the person selected is deleted from the database to prevent duplication. The following describes the discussion page. When you enter this page, the donated books are displayed on the screen. Besides, the information of those who apply anonymously is shown e-mail instead of their names. This page stores the entered information, such as gift pages, in the database and shows the address to which to send the donation book to the user when the application is complete. Finally, on the recommendation page, users can view the books recommended by other users by genre, and see a line of reviews. A single-line review form creates a book name, author, and one-line review, which is stored in the database. Reviews that have been created can be viewed on the page. It also limits the number of reviews to reduce memory waste, so that when the number of reviews exceeds a certain number, the oldest reviews are deleted.

4 Implementation

This section shows specific implementation results of web pages.

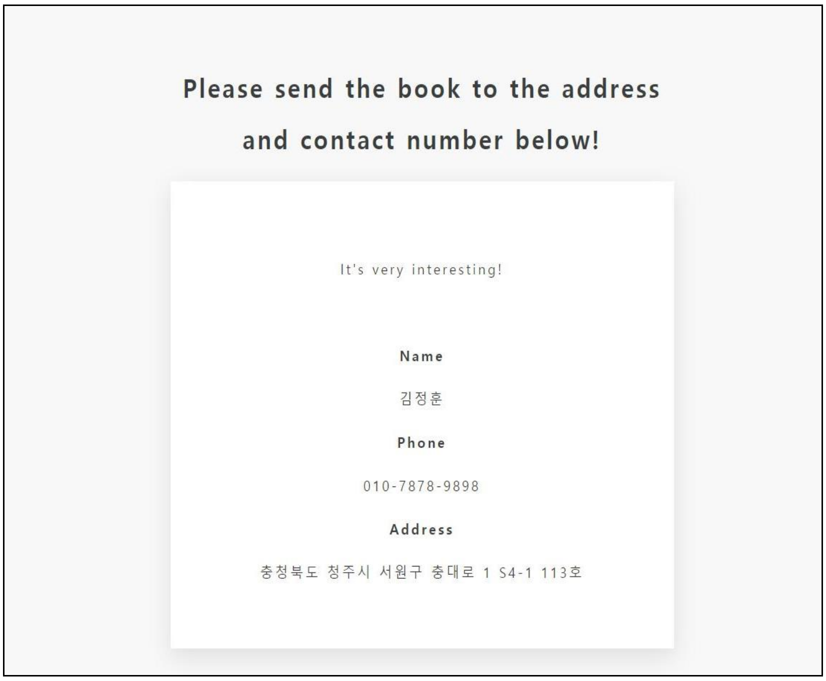


Welcome to Book Exchange Project!

Application

Please fill out the book for the person who will present it to you.

Figure 3 Application form in Gift page



Please send the book to the address

and contact number below!

It's very interesting!

Name

김정훈

Phone

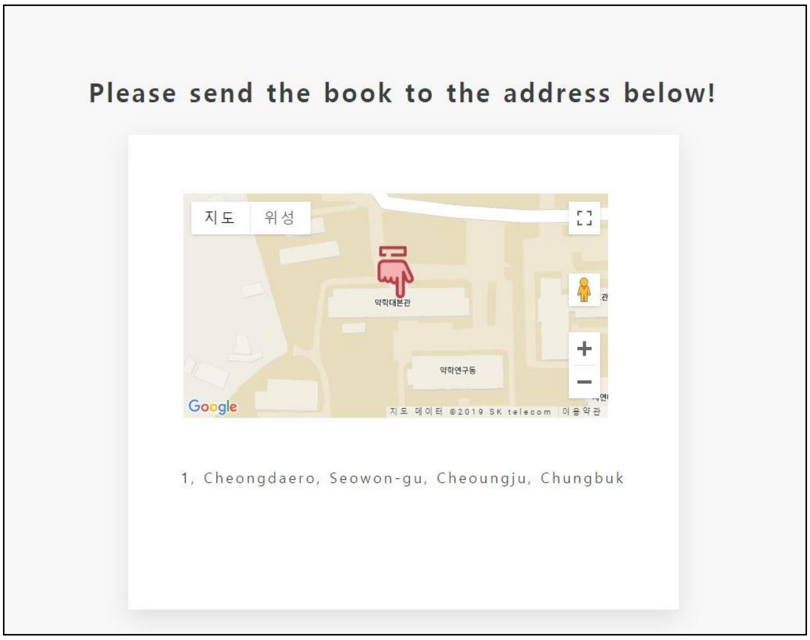
010-7878-9898

Address

충청북도 청주시 서원구 충대로 1 S4-1 113호

Figure 4 Application Result in Gift page

Figure 3, 4, is the Gift page. When the applicant fills out the application form, information to randomly send to another person is displayed. In other words, the stored data is immediately displayed on the screen, except for the applicant's information.

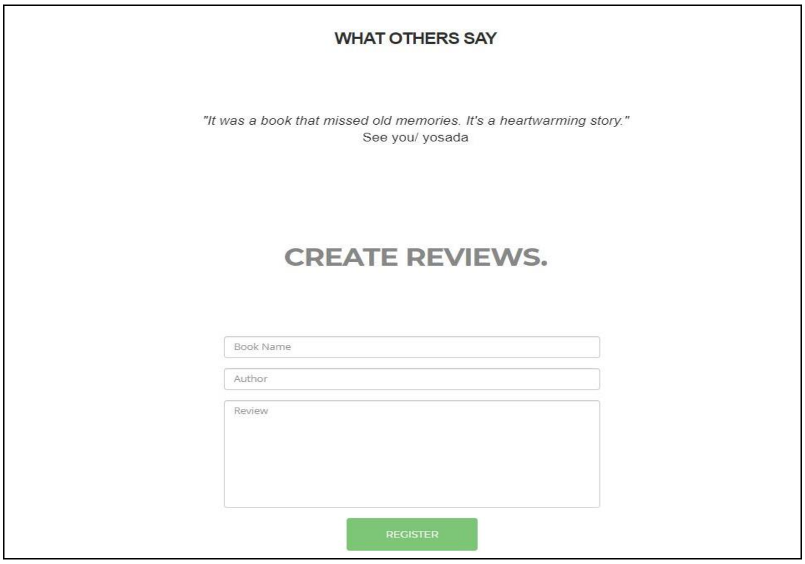


Please send the book to the address below!

1, Cheongdaero, Seowon-gu, Cheoungju, Chungbuk

Figure 5 Application result in Donation page

Figure 5 is the Donation page. When the donation application is completed, show the user the address to be delivered using the Google map API.


WHAT OTHERS SAY

"It was a book that missed old memories. It's a heartwarming story."
See you/ yosada

CREATE REVIEWS.

Figure 6 Application form & Review slide in Recommendation page

Figure 6 is the Recommendation page. On this page, a list of recommended books is posted at the top. At the bottom of the page, a registered review is shown in slide format, as shown in Figure 7. There is also a form with which you can write a single line review. All pages used JAVASCRIPT and JQuery languages based on HTML and implemented slide and scroll animations using Query plug-in. Besides, the main page was applied to become a reactive web page by using the template in the bootstrap.

5 Conclusion

Books provide intellectual nourishment, as we have always said, is the best way to enhance our thinking and judgment. However, with the widespread use of smartphones and the Internet, the reality is that children and adolescents are growing up without proper reading habits. The effect of reading is sufficiently correlated with reading in research in various fields. You can also see that reading is related to the learning performance of university students, the Korean language performance of teenagers, and the creative thinking of infants. There are many ways to engage in reading activities in relation to studies during the school year, but this only negatively affects the perception of reading and makes it difficult to get interested in reading on one's own.

As I said earlier, the reading volume is so low that the nation is also preparing measures to increase it. However, as the development of information and communication technology increases exposure to the media, people are bound to be drawn to more stimulating, faster, and easier image media. To this end, we have come up with a book exchange project that is just as interesting. This can stimulate people's curiosity to become closer to books, and naturally tame reading habits. This Web service creates an opportunity to get closer to a book, starting with finding a book to send as a gift in my home, in my room. There are many related events in university libraries or local libraries, but it is the most difficult thing stepping to there, so people can use the web services more easily they use a lot these days.

Reference

1. Yoon, Sang-Ho. “The economic impact of reading.” Korea Economic Research Institute. (2016. 5)
2. Lee, Kyung-Min. “A Study of Reading Education Methods for University Students in University Libraries.” Korean Library . journal of information society, Book 43, No.4 (2012. 12)
3. Kim, Eun-Joo. “An analysis of the structural relationship among professor-students interaction, class satisfaction, reading and learning outcomes of students.” Comprehensive Education Research, 12:3, 1-22 (2014)
4. YTN SCIENCE Web, <https://science.ytn.co.kr>

Work in Progress: AI World Cup Competition

Donghyeok Lee[1], Jeongeun Ahn[2], Choyoung Kim[3]

[1] Department of Commerce and Trade, [2] Division of Computer Convergence,
Chungnam National University, Daejeon 34134, Korea
[3] Department of Comuter Science & Engineering, Pusan National University, Korea
{24606937a, lacri9159, kimchoyoung7}@gmail.com

Abstract. This paper discusses how Artificial Intelligence (AI) is implemented to AI soccer, one of the AI World Cup competitions. We analyze the strategies of the actual soccer competitions and the previous AI world cup competitions. The AI World Cup competition was once held, so there is very little data available. We applied two methods to learning machine: logic and Q learning. Three top priorities of the game strategies have been presented. This preliminary results will be revised and verified until the 2019 AI World Cup.

Keywords: AI World cup, strategy, logic, Q-learning

1 Introduction

Artificial intelligence (AI), which is one of the technologies that lead the 4th industrial revolution, has become big attention. And AI is rapidly applied to real life. AI, such as recommendation system, face recognition technology, speech recognition, is used in various places. The game field also uses AI, which includes personalized content and game data analysis. Korea's top Go player Lee Sedol, and Google's AI AlphaGo played Go matches in 2016. AlphaGo won the match and surpassed all expectations [1]. Interest in AI has increased among the public, research on AI has become active, and the field has expanded. Interestingly, the AI World Cup was held for the first time in KAIST as a sports game using AI. The AI World Cup Competition was the only competition in which participants from around the world competed using a learning model in soccer games [2].

In this paper, we discuss how AI applies to AI world cup competition. It focuses on approaches and strategies rather than a detailed description of the implementation code.

2 Related Works

AI World Cup is the world-wide competition of online soccer teams simulated by AI. The AI World Cup aims to inspire the future of AI technology through innovative solutions and technological competence. The scope of the AI World Cup includes AI soccer, AI commentator, and AI reporter. Among these three challenges, this paper discusses AI soccer. Five AI soccer players, controlled by AI, move around the field to kick the ball into the opponent's goalpost [3].

Webots is an open-source robot simulation software. It provides an environment for robot modeling, programming and simulation. In here, it provides an environment where robots can play soccer virtually.

Q learning is trained by Q function, which requires state, action as input and provides expected Q values. The Q function tends to choose the action that returns back the highest Q value. By repeating this process, it selects the best policy in provided environment. One of the advantages of Q learning is model-free algorithm [4, 5].

3 Methods

In the 2018 FIFA World Cup match between Korea and Germany, Germany used strategies to put pressure on the front and to target the opponent's weak point with the edge. They also keep their own positions well and have a strategy to change the offensive and defensive positions. In last year's AI World Cup competition, teams that use an unconditional aggressive strategy scored higher and some teams without goalkeepers attacked all the players. It also showed a strategy of marking the opponents without following the ball [6]. The top priorities of the strategies were summarized based on the above game analysis.

- Step 1: Kick the ball right after the games starts toward the opponent's goalpost.
- Step 2: Get into position and identify which side of the ball is in. If it is on our side, the defense starts. If it is on opposing side, continue attacking.
- Step 3: Attack and defend on the other side of the center line.

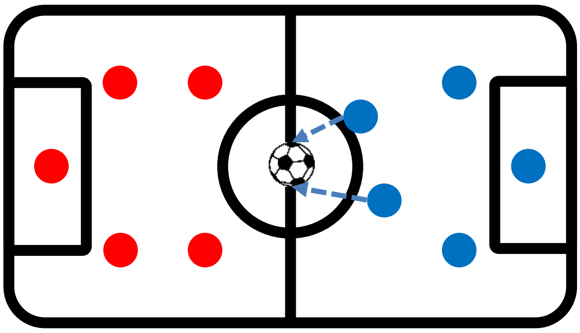


Fig. 1. Each team has five players, and each player's role can be set differently.

This AI World Cup competition has not been held many times and the game is not public, so there is very little data. Therefore, we decided to use logic and Q learning.

3.1 Logic

There are 11 players per team in a real soccer game, but in this competition there are 5 players, including a goalkeeper for each team (shown in Fig. 1). It is important that all players finish the game without being sent off. There is no need to have a defender in the opponent's area where only the goalkeeper is left. Therefore, strikers, defenders and goalkeepers are flexibly divided according to the number of players left in each team. The number of players per position depends on the position of the ball. After measuring the distance of the players, we arrange them in ascending order. Based on the distance between the ball and the player, the nearest players are placed in order of striker, defender, and goalkeeper.

3.2 Q Learning

It is essential to set the reward of the q-function in Q Learning. If the robot has a ball, the robot has to own the ball and run to opponent's goalpost. If the robot does not have a ball, the robot should get closer to the ball. When the frame representing the game screen is changed, an action is determined based on the position of each robot. In terms of reward, when the robot does not have a ball, it is set to give the reward based on the distance of the robot from the ball. When the robot has a ball, it is set to give out the reward based on the distance from the goal area of the opponent.

4 Preliminary results

This paper discussed approaches and strategies for how AI applies to AI soccer, the field of AI world cup. The content of the previous competition is not public, so there is less data. Therefore, logic and Q learning were used in this study. In our logic, the distance of the players is measured, and the distance between the ball and each player determined their role. Using the Q learning, each reward was set according to whether the robot has a ball or not.

However, the proposed AI soccer strategy has not yet been fully implemented. There was a lack of concrete strategies considering various situations. It took a long time to implement and verify test results to ensure if the strategy worked. It seems that more situations need to be considered until the 2019 AI World Cup Competition, and a better strategy is expected based on these.

Acknowledgment

This research was supported by the Korean MSIT (Ministry of Science and ICT), under the National Program for Excellence in SW (2015-0-00930), supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Jo, B. H., Park, C. J.: Research Trends in Game AI, Electronics and Telecommunications Trends, 23(4): pp115-121 (2008) (in Korean)
2. AI world cup Competition, https://github.com/aiwc/test_world
3. AI World cup 2019 official Website, https://aiworldcup.org
4. Deep Q Network, https://poqw.github.io/DQN/
5. Yang, D.: A Study of Tennis Game AI through DQN, Thesis, Korea University (2018)
6. 2017 AI World cup Final (WISRL vs AR LAB), https://www.youtube.com/watch?v=bi9yT1XhSgI

Development of Custom Phone Case Shopping Mall Using Web Open Source Software

Changhyun Kim, Jiyoung Oh

Department of Computer Science, Chungbuk National University, Republic of Korea
{ckdgus0204, y7133}@cbnu.ac.kr

Abstract. Currently, all ages of smartphone users are increasing, and the size of the smartphone accessories market is overgrowing as well. However, in the era of individuality, the accessory market that meets all user needs is still small. By improving sites that offer a similar design, the project allows users to design smartphone cases directly to meet their needs by taking advantage of their unique personality. So, the project can increase the user's choice, and It increases the level of satisfaction in purchasing and contributes to the expansion of market size.

Keywords: web technologies, customize, open source software

1 Introduction

As smartphones became available worldwide, the number of people buying smartphone accessories also increased exponentially. As types of Smartphones increases and amount of use increases, markets for accessories for Smartphones are also expanding. According to an analysis of actual consumer usage patterns, KT Economic Management is estimating that the size of accessory markets will amount to about $1.6 trillion (1.6 trillion KRW). As the number of new entrants to the market surged, the accessory market also grew at an annual average rate of 62 percent.
Of course, it is the smartphone case that accounts for the highest percentage of smartphone accessories. Those who purchased smartphones not only in their 20s and 30s but also in their 50s and 60s own more than one smartphone case.

Have it in possession. Why do people buy not only smartphones but also smartphone cases? The front screen of the smartphone uses glass material, and the cover on the back also uses plastic or glass material. Scratches and shocks can easily break plastic and glass. Smartphone cases cover easily damaged Smartphones, and the cover glass is also protected. As such, smartphone cases have the advantages of safeguarding smartphones from scratches and shocks. According to a U&A survey on smartphone case purchases conducted by Macromillembrain, 91 percent of the respondents said the reason for buying smartphone cases was to protect their smartphones from "scratch or damage."

Besides, 66.8% of future intent to buy a smartphone case was high, and the main factor of consideration for re-buying was a smartphone protect as well and external beauty. Adequate smartphone protection of smartphone cases has been the driving force behind the repurchase. External aesthetic factors, another driving force behind the repurchase, tell us that many people prefer products that show beauty and individuality. Of course, the cases already made will be pretty, but user-created cases will be more meaningful to users and closer to their taste. Most smartphone case shopping malls, however, cellphone cases that users have already made, not created.

We created a smartphone case shopping mall where users can decorate their phone cases according to their tastes with smartphone cases, a typical item that protects smartphones.

2 Research Objective and Goals

We wanted to acquire the whole process from Web site development to distribution. The goal is learning through actual development rather than theoretical learning. Among the various Opensource software that has become a recent issue, We would like to use various functions and use a range of Django, Nginx, uWSGI, and Google maps API related to the web. We want to create a dynamic web that can communicate with a server using Django, a developed framework on a static web using only Html, CSS, and JavaScript, and store data. Besides, since the main customers are those who want to make phone cases, we create a Responsive web to build an optimized environment that is easy for users to use. By connecting various other sites as users and finding out if there are any inconveniences, we will develop a web page that has a difference from existing sites by applying the advantages of using it. To facilitate maintenance, later on, each code is separated by use and annotated to improve understanding.

3 Function Analysis of Phone Case Shopping Mall

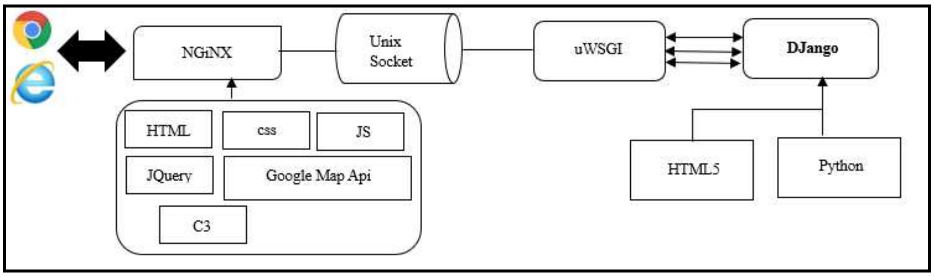


Figure 1 Flow of the web page

Figure 1 is the working flow until the open source used by the webpage is expressed in the web browser.

The project used Django, a full-stack web framework written in HTML5 and Python. A Web client sends a request to the Web server to the HTTP protocol.

Static files can be processed directly from the web server NGINX.
Dynamic files delegate to web application server uWSGI because the web server cannot process them. The web server delegates the request to the web application server, and the web application server responds to the web server on its behalf. The Unix Socket interacts between these.
Django framework executes application code written by the user in Python. The design was designed using HTML and CSS to decorate the interface, and the functional parts were JavaScript, JQuery, C3, Google MAP API. Users can view the web pages of this project using Internet Explorer or Chrome.

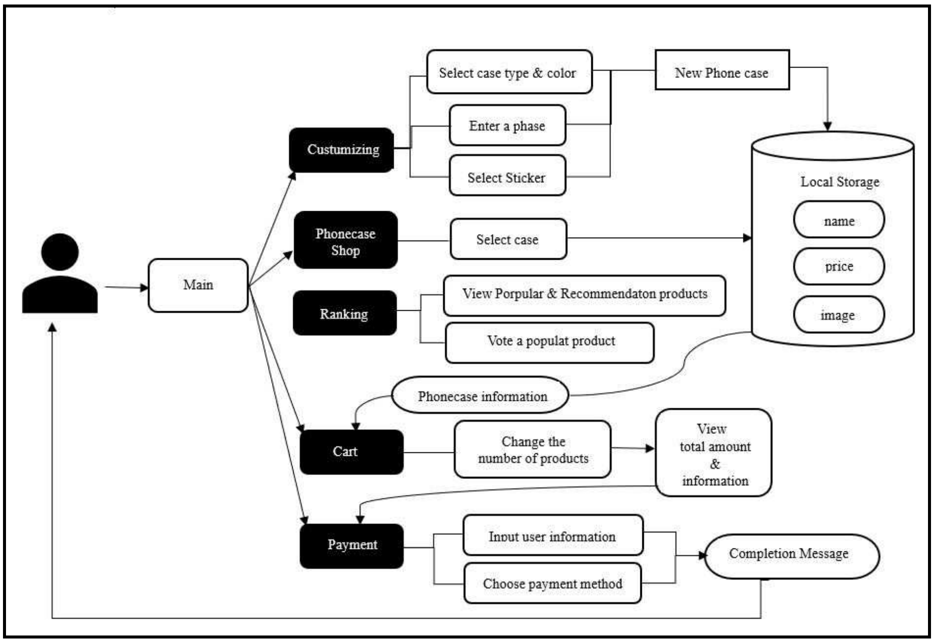


Figure 2 functions of structure

Figure 2 shows the pages and functions of the website. It also shows the flow of connections between pages.

3.1 Main Page

The top shows a beautiful and colorful phone case images used sliders. Touch four buttons of the center to go to that site. And at the bottom, the map is displayed in Korea using Google MAP API, and the location of the shopping mall company is shown through the latitude and longitude information of the shopping mall. The statistics next to the map were C3 using the Simple XY Line Chart, the x-axis shows a week, and the y-axis shows people in 1 unit. A graph of a broken line of each male and female shows how many people came to the site a week.

3.2 Customizing

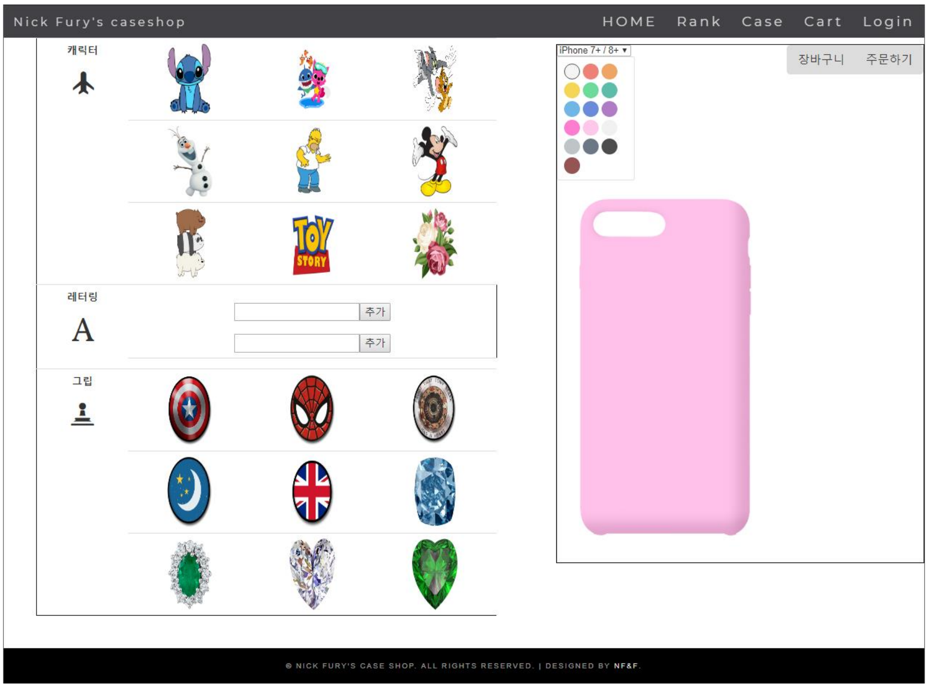


Figure 3 Customize page

Figure 3 is a page that allows users to customize the case they want. You can select the elements you want to add and decorate them on the chosen case.

When the user selects the model and color from the select tag, it uses document.getelementById to get the model name and color. It then outputs the image by replacing the corresponding .src element with the corresponding image name and color name. On the left side of the screen, you can use JavaScript to output the elements you can add. You can click on the item you want to add and drop it onto the case.

```
program drag (e)
  var  eventobject: window.event?window.event:e;
  if(eventobject.className==”drag”)
    this.offsetx=eventobject.left
    this.offsety=eventobject.top
    this.x=eventobject.clientX
    this.y=eventobject.clientY


program letter()
  if(document.getElementById(“change”+i).value==i)
  var txt=document.getElementsByClassName(text)[i].value
```

```
document.getElementByName(text)[i].innerHTML=txt
end.
```

The drag function is a function that is executed when an image is pressed. If it is an explorer, replace it with window.event because the event object e does not exist. Since drag and drop should be able to move only the decorative elements, only the class with the dragged name of the class name of the decorative element is made movable. Dragging and dropping is performed by pressing the mouse while calculating the distance of movement of the image using offset x and offset y, and changing the position of the image accordingly. Move the object's full-screen position using the object's clientX and clientY. When using e.preventDefault, the most important drag and drop functions are executed by using the function that stops the function of the tag when clicked and preferentially follows the function of JavaScript. And if you want to add text, write a text in the text entry and click the Add button to create and add the text. Once you have added all the elements you want, click on the shopping cart and use the canvas API to capture the phone case section image. And add the corresponding charge to your shopping cart using local storage.

3.3 Phone Case Shop

It's a collection of pre-manufactured products. It shows three case products: Jelly Case, Slim Hard Case, and Buffer Case. These three cases created used a constructor and showed the case image, price, and name using an array of 16 corresponding cases within each case. If the user wants to put the case in the user's shopping cart, the user can also cancel it by selecting the shopping cart button below and pressing it again using the 'mouseUP'.

If a user selects the shopping cart button, the user can go directly to the shopping cart page or continue shopping.

3.4 Ranking

The ranking shows the image, name, and price of a popular product. User can press the 'Good' image button at the bottom to change the number of 'Good'. Recommended products appear on the bottom, and if the user clicks on the name, the user will be transferred to the phone case shop page.

3.5 Cart

The shopping carts list items that user-customized case or selected from existing cases. This is shown as an array using local storage. User can change the number of selected products and the price changes accordingly. Suggested products appear on the bottom, and if the user clicks on the name, the user will be transferred to the phone case shopping page.

Once the user has checked your shopping cart, the user can go to the payment page through the button.

3.6 Payment

User can enter the user's name, number, and address. The system Indicates the total amount of user's delivery items, including shipping amount. And choose a method of deposit between account-free deposit or credit card payment. After selecting, a message appears through the 'alert' that the payment has been made.

4 Design of webpage

In the case of an existing website, the screen often appears differently depending on the size of the user's screen. As a result, users may feel uncomfortable and sometimes unable to use the desired functions appropriately. So my design has been designed to responsive web. As smart devices evolve into various forms, users do not use the same proportions or screen sizes, minimizing user inconvenience by providing optimized screens for each screen. When you reduce or increase the size, it is scaled down to provide an expanded view. And various designs were applied using bootstrap. Pressing a button or hovering the cursor will make the user more familiar with the visual aspects of the interface, layout, and structure in response to the user's familiarity.

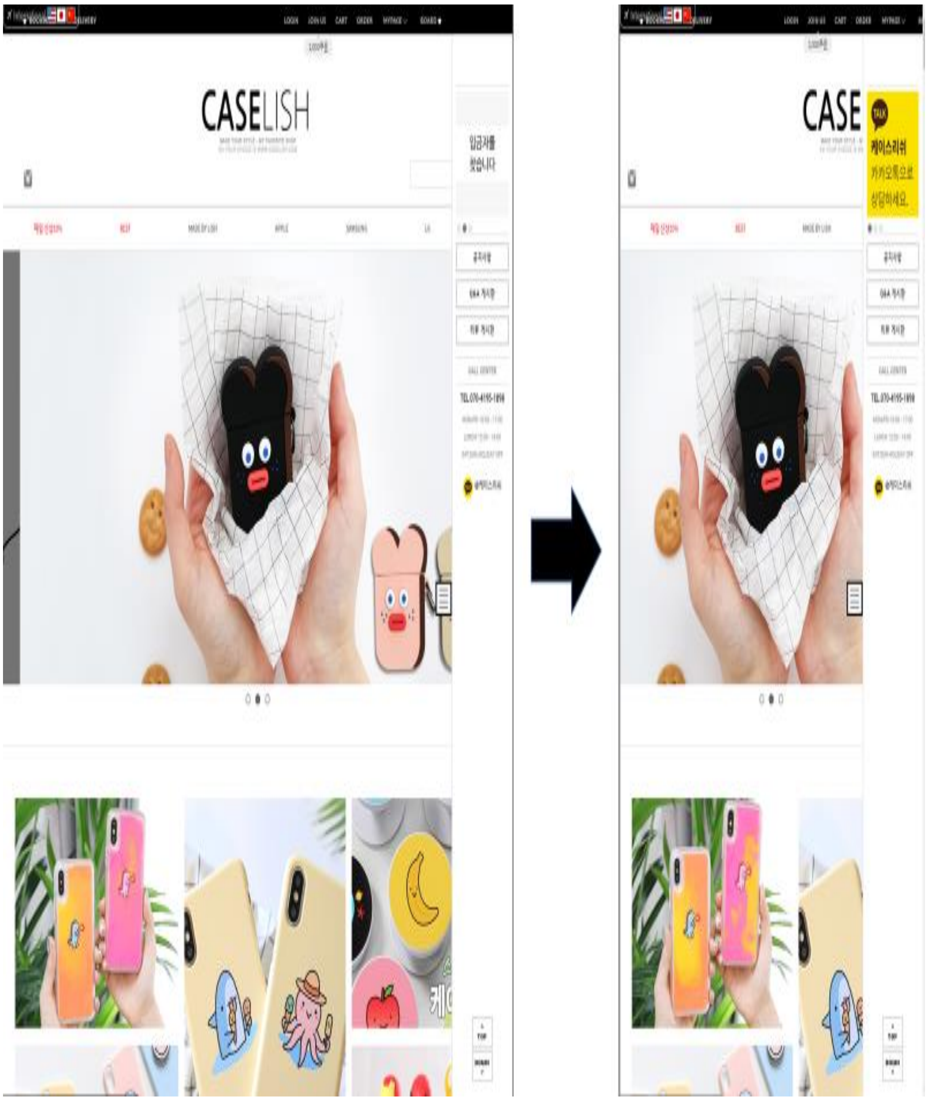


Figure 4 Other web pages not a responsive web

Figure 4 is a website that does not have a reactive web. Some things are difficult to use because they do not decrease in size.

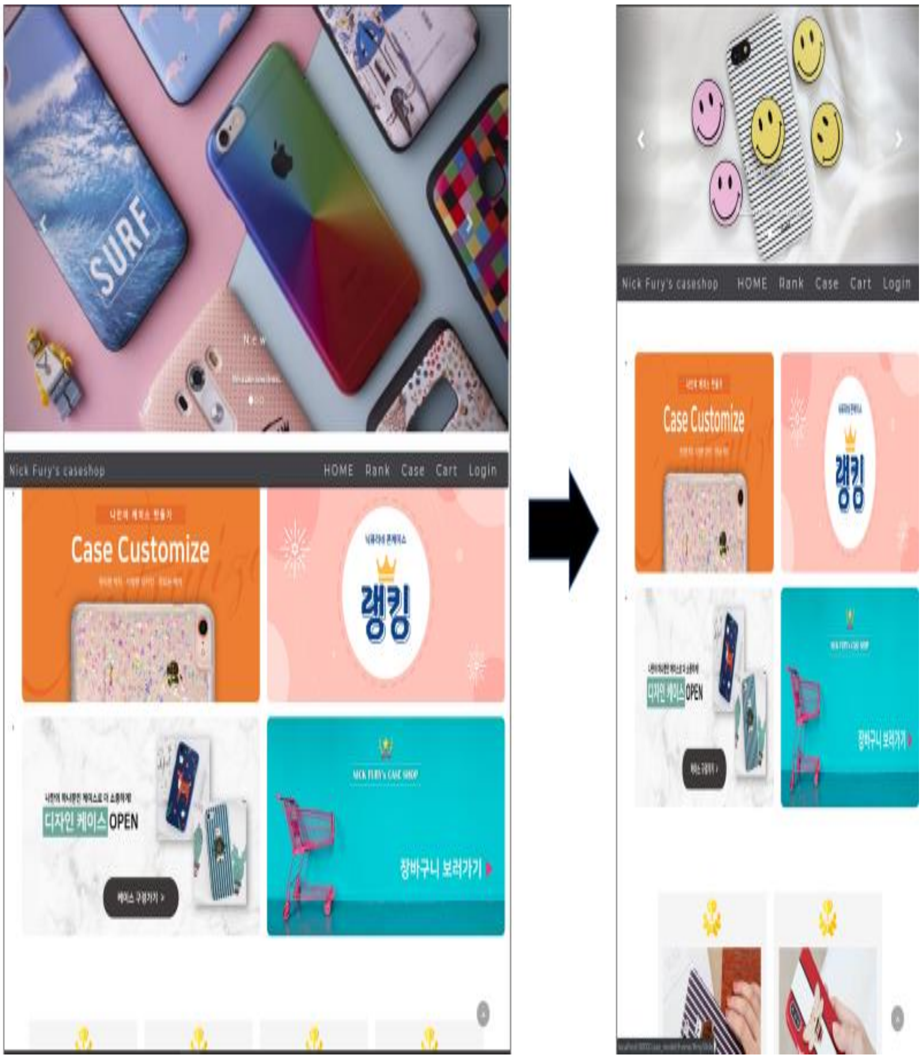


Figure 5 responsive design web page

Figure 5 is a website with a reactive web. When reduced in size, the components are reduced in size together, allowing them to perform their original functions.

5 Conclusion

This project is a shopping mall of phone case customizing that can be produced directly. Users can choose a pre-built case as well as be self-produced. Users can select a product and put it in a shopping cart to purchase the product.

And with a ranking page, you can see which is a popular product, and the user can adjust the ranking of the product by pressing the button of Good. It also uses Google

Map API to show statistics on the location of shopping malls and how many users came in through C3.

These features can increase the user's choice, and It increases the level of satisfaction in purchasing and contributes to the expansion of the market size.

References

1. HoYeon Yang, ‘phone case market is growing up’,2013.05.16 ,https://it.donga.com/14491/
2. YoungHun Song, WonGuen Hong, 2013.04.24. “going to be major, phone cases”, DigiEco Report, p.8
3. Moore, FahmeenaOdetta. (2016). Responsive Web Design. 10.13140/RG.2.1.1555.4160.
4. Rubio, Daniel. (2017). Introduction to the Django Framework. 10.1007/978-1-4842-2787-9_1.
5. Editorial department. (2012). U&A research about Smart phone case purchase 12(1), 2-28.

Design and Implementation of Unmanned Restaurant System using HARMS

Chihyeon Ryu[1], Seungju Yoo[2], Dongho Nam[1], Yunsu Jung[1], Seungjin Baek[1], Minsun Lee[1]

[1] Division of Computer Convergence, [2] Department of Linquistics,
Chungnam National University, Daejeon 34134, Korea
{rch8952, seung3389, dongho0916, haon4658, bsj1048}@naver.com, mleeoh@cnu.ac.kr

Abstract. Utilization of the unmanned system through robots in food-distribution businesses maximizes customer convenience and reduces labor costs. To build ubiquitous systems through robots, agents must communicate with each other. This way, each agent can perform given tasks by exchanging messages. In this paper, we present the Unmanned Restaurant System based on the Human, Agent, Robot, Machine, and Sensor (HARMS) and implement the seat selection model using three agents. Proposed system exchanges messages between the agents to process given situations. The results show that the configuration of the ubiquitous environment pursued by HARMS was successful.

Keywords: unmanned system: multi-agent system; HARMS

1 Introduction

As robotic engineering develops, automated systems using robots are replacing various jobs that humans have been doing [1]. The establishment of the unmanned system using robots increases the productivity of society as a whole and enables humans to concentrate on higher-level tasks. The Kiosk is one of these automated systems [2]. It is an unmanned ordering system that enables customers to search, order and make payments themselves. The kiosks that could only be found in movie theaters and airports are being introduced into a variety of industries, starting with fast food stores. Eighty percent of the users said that the advantage of kiosks was that they did not have to wait in line or make unnecessary communicate [3, 4].

The fact that simple labor is replaced by automation systems means that employees can concentrate on providing high-level services through mutual exchange with customers. However, current automation systems can only respond to customers' unilateral demands. It cannot completely replace human jobs because the automation system cannot interact with the customers and cannot offer better choices.

Through the robot system ubiquitous, HARMS, which stands for humans, software agents, robots, machines, and sensors, aims to maximize the intervention (substitution ability) of robots in human life, so that humans can perform higher-level tasks [5]. With such goal, HARMS is designed implement ubiquitous, which includes humans, to replace human jobs. The communication system of HARMS, done through interactions between agents, is enough to substitute humans not only because it can provide efficient labor and convenience of customers, but can also provide customized services.

In this paper, we have designed an automated system restaurant based on HARMS to replace employees (human). Inside the HARMS ubiquitous, humans and agents continuously interact with each other. The procedure of the agent responding to a human's demand and finding the right agreement is one of the interactions. The system using HARMS has the advantage that it can be easily implemented in various environments using one ubiquitous database. We implement the seat selection model in the restaurant using three agents: Table Agent, Receptionist Agent, and Customer Agent.

The remainder of this paper is organized as follows. Section II outlines the Unmanned Restaurant System in detail. Section III describes the implementation of the system and the results. Finally, Section IV presents conclusions and discusses future work.

2 Unmanned Restaurant System

The HARMS system, which is the basis of Unmanned Restaurant System of this paper, constitutes the ubiquitous environment [6, 7]. It consists of five abstract layers (Network, Communication, Interaction, Organization, and Collective Intelligence. The Network layer represents the inter-actor network connections of the HARMS system. We focus on wireless network connection between actors and enables unicast, multicast, and broadcast message transmission. The Communication layer defines the protocol for communication between actors and defines the syntax for message exchange. The protocol, that defines the message number as the context request, is constructed and a GUI interface is provided for the communications between actors. The Interaction layer represents algorithms and techniques for rational decision making of actors in inter-actor communication. We use an algorithm to identify and exclude objects based on the requested message. The Organization layer organizes aggregate intelligence by gathering inter-actor decision-making across the three tiers of network, communication, and interaction. The Collective Intelligence layer enables you to reach optimal results in the fastest time possible with the collective intelligence you have configured.

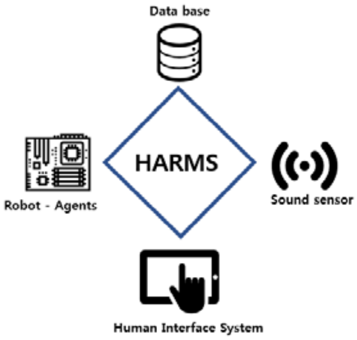


Fig. 1. Interaction of HARMS Agents diagram

Proposed System based on HARMS operates through mutual communication between all agents, shown in Fig. 2. The operation process according to time is as follows. First, the operation of this system starts by sending a message to "Receptionist Agent" via start button of "Customer Agent". When a customer enters a restaurant, they can choose their own location through the "Customer Agent", which works as a tablet. Customers will be recommended for seating based on the number of people, drinking, and noise levels. Once assigned, the customer will be able to select and order the desired menu via the "Customer Agent". The customer's order is delivered to the "Kitchen Agent". When the ordered menu is completed, "Kitchen Agent" requests the "Waiter Agent" to serve the order. After the customer leaves the table, the "Waiter Agent" sends a request message to the "Table Agent" to clear the occupied state of the table. The "Table Agent" that receives the message converts its own table occupation state into the occupancy states.

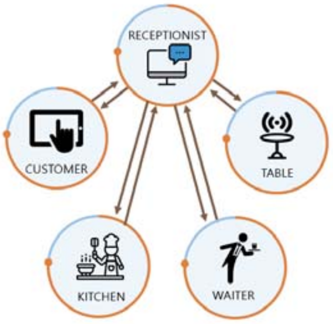


Fig. 2. Message exchanges between agents

3 Implementation and Results

Customer Agent is a HARMS-based mobile application. It is designed based on JAVA language and serves as a mediator to communicate with other agents by providing GUI to customers. In addition, it intuitively visualizes messages received from other agents. Customer communicates with "Receptionist Agent" through "Customer agent". When the application is executed, the message is delivered to the "Receptionist agent" and the peer (agent) is added to the communication list automatically so that the communication can be continuously performed. The "Customer Agent" is designed to ask the customer for a placement option for the number of people, drinking, and noise levels and receiving a response from the customer. After receiving the response to the query to the "Receptionist agent" and receiving a message that the customer has a desired seat, the next question is passed to the following activity. On the other hand, if you get a message saying that you do not have an available seat, customer agent suggests the customer a next best solution alternatively. This entails artificial intelligence technology.

Receptionist Agent is a server agent based on HARMS system. Receptionist Agent is designed based on JAVA language and has table list information for storing table responding to customer's seat selection requirement through Customer Agent. The Receptionist Agent acts as a mediator between the Customer Agent and the Table Agent to select the customer's desired location. Through the Customer Agent, the customer sends a message to the Receptionist Agent with three levels of information: number of people, drinking status, and noise level. The Receptionist Agent sends a message to the Table Agent in a broadcasting and multicasting manner at every step to check the table satisfying each step, and records the response from the Table Agent satisfying the requirements in the table list. Each time you perform each step, the number of tables that need to be examined is reduced. Finally, the table number of the customer-selectable table is sent to the Customer Agent as a message.

Table Agent is an agent that operates on Raspberry Pie based on HARMS system. Table Agent is designed based on JAVA language. It is composed of table number to identify each table, number of seats that table can accommodate, availability of drinking in each table, noise level for each table position, and occupancy status information. The attribute values of each of these tables are used through mutual message exchange with the Receptionist Agent and the Waiter Agent, allowing the customer to select the desired seat and show the occupancy status of the table by the customer. The Customer Agent delivers the customer's seat selection requirement information to the Receptionist Agent through the message. The Receptionist Agent then transmits a broadcasting message to the Table Agent, and the Table Agent replies with the table number that satisfies the customers' demands.

We implemented the customer positioning from the exchange of messages between Customer Agent, Receptionist Agent, and Table Agent in the proposed system. This test is largely divided into implementation of three objects: Customer Agent, Receptionist Agent, and Table Agent.

- Customer Agent is implemented with tablet applications running on Android OS version 7.0.
- Receptionist Agent and Table Agent operate on the Raspberry Pie 3 running the Raspbian OS and provide a GUI through the Java Swing toolkit.
- Table Agent uses a separate sound card [1] and microphone for noise measurement.

In order to communicate between agents, numberings according to contextual messages was inevitable. Table 1 shows the assignment of message numbers according to message type and communication flow between agents. The first and third digits of the message number are defined as the number of the communicating agent and indicate the direction of the communication. The second digit of the message number has been assigned from 0, depending on the stage at which the message is sent.

The message transmission between agents is done through Java socket communication. Before communicating, each agent is implemented to be ready for communication by adding the agent to communicate to the "peer list". When the program is executed, each agent automatically adds peers to the program accordingly from the list of peers created. Press the "Start" button on the Tablet GUI of Customer Agent to start operation of this system. Each agent receives the corresponding message number through the Listener method, and performs and calculates the requested work as the internal logic of the function. After sending the return value (table number,

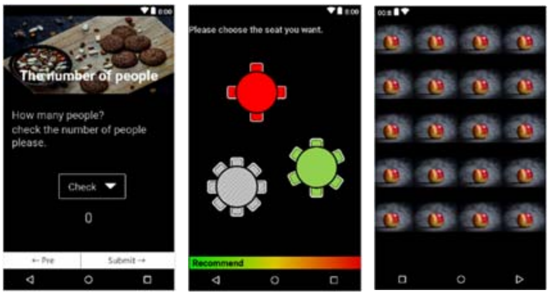
[1] Stereo Raspberry pi sound card, Audio Injector, http://www.audioinjector.net/

number of guests, flag to go to the next stage) in message, the Customer Agent is implemented so that it can move to the next stage after receiving the message of allowing the receptionist agent to proceed next stage.

Table 1. Message protocol of Unmanned Restaurant System*

Stage	Sending Agent	Receiving Agent	Type	Message Number
Start	Customer	Receptionist	Unicast	102
	Receptionist	Customer	Unicast	201
Number of people	Customer	Receptionist	Unicast	112
	Receptionist	Table	Broadcast	213
	Table	Receptionist	Unicast	312
	Receptionist	Customer	Unicast	211
Seat assignment	Customer	Receptionist	Unicast	142
	Receptionist	Table	Unicast	243
	Receptionist	Customer	Unicast	241

The system includes actions of seating arrangements, order, and the releasement of the seat after the customer leaves. Fig. 3 shows the three stages of Customer Agent from the start to selecting menu. In this paper, we implemented the seat selection model using three agents: Table Agent, Receptionist Agent, and Customer Agent, excluding Waiter Agent and Kitchen Agent. Therefore, in order to implement a complete system, we need to extend the scenario for specific actions, including the remaining two agents.



The number of people

How many people?
check the number of people please.

Please choose the seat you want.

Fig. 3. Three stages of Customer Agent

4 Conclusions

The significance of the HARMS system is to replace human work by building a ubiquitous environment using robots. The system of this paper exchanges messages between the agents to process given situations without any problems. As a result, configuration of the ubiquitous environment pursued by HARMS was successful. Each agent successfully communicates as the message protocol set for each situation and functions correctly. The HARMS system proved to be able to build a strong ubiquitous environment in a multi-agent system.

Proposed system includes actions of the seating arrangements, order, and the releasement of the seat after the customer leaves. This system makes it easy to implement additional functions as well as measuring the noise during customer assignment. It can be expanded just by adding necessary agents according to the environment of the restaurant, so it has high scalability and flexibility.

Among the five layers of HARMS, the Organization and Collective Intelligence layers have not been implemented, and the addition of necessary artificial intelligence elements in the decision-making process between agents is needed. Afterward an additional task that has to be extended to this system is to recommend the best place for the customer who does not have the desired place, based on the input data received by the agent.

Acknowledgment

This research was supported by the Korean MSIT(Ministry of Science and ICT), under the National Program for Excellence in SW(2015-0-00930), supervised by the IITP(Institute for Information & communications Technology Promotion)

References

1. Royakkers, L., van Est, R.: A Literature Review on New Robotics: Automation from Love to War. : Int J. of Soc Robotics (2015) 7: 549. https://doi.org/10.1007/s12369-015-0295-x
2. Hofelder, W., Hehmann, D.: A Networked Multimedia Retrieval Management System for Distributed Kiosk Applications.: 1994 Int. Conf. on Multimedia Computing and Systems, Boston, USA, IEEE Computer Society Press, pp. 342-351, (1994)
3. Garnick, C.: Costco's testing self-service food kiosks in a Seattle-area warehouse, https://www.bizjournals.com/seattle/news/2018/02/21/costco-wholesale-self-service-food-kiosk-tests.html (2018)
4. OLEA Kiosk: The Demand for Retail Self-Service Kiosks Is on the Rise!, https://www.olea.com/thelab/the-demand-for-retail-self-service-kiosks-is-on-the-rise/ (2015)
5. Lim, H., Kang, Y., Lee, J., Kim, J. and You, B.: Software architecture and task definition of a multiple humanoid cooperative control system.: Int. J. Human. Robotics, 6(2), 173-203, (2009), https://doi.org/10.1142/S0219843609001747
6. Matson, E. T., Taylor, J., Raskin, V., Min, B., Wilson, E. C.: A natural language exchange model for enabling human, agent, robot and machine interaction. In: The 5th International Conference on Automation, Robotics and Applications, Wellington, pp. 340-345, (2011) doi: 10.1109/ICARA.2011.6144906
7. Kim, M., Koh, I., Jeon, H., Choi, J., Min, B., Cho, I., Matson, E. T., Gallagher, J.: A HARMS-based Heterogeneous Human-Robot Team for Gathering and Collection Function. Advances in Robotics Research, Volume 2, Number 3, pages 201-217 (2018)

A Web Open Source Based Health Monitoring System for Controlling Eating Habits

YongGee Kim*, WanSeok Lee*, YeonSu Kang*
e-mail : ykkim6872@gmail.com ,k1nder@naver.com, rkddustn96@naver.com

*Department of Computer Science, Chungbuk National University, South KoreaSpringer-

Abstract. Recently, the proportion of single-person households across all ages has been on the rise. The reality is that one-person households are forced to live alone, showing a wrong diet. In order to solve this problem, it is deemed necessary to promote a healthy culture and dietary habits of single-person households through media with reasonable accessibility. Therefore, this paper aims to provide not only a popular food culture, but also a good recipe for reading, marking the expiration date of materials, checking information on nearby markets using GPS, and chatting bot for easy cooking. This is to help improve eating habits for many college students who live on their own, and even for single-person households.

Keywords: We would like to encourage you to list your keywords in this section.

1 Introduction

As the aging society progresses rapidly, the number of older people living alone is increasing; also the proportion of single-person households by all ages is growing due to the lack of marriage and low birth rates among those in their 20s and 30s at the right time of marriage due to the shrinking population and economic burden.[1]

As single-person households have to live alone, they have many problems compared to multi-person households in eating habits and food culture. Of the nutrients, dietary fiber intake was the lowest, especially among households aged 19-29 compared to multi-family households. The lower the age, the higher the rate of breakfast, the higher the rate of breakfast for a single-person household compared to a multi-person household in all age groups, and the higher the frequency of consumption of carbonated drinks, instant noodles and fast food such as hamburgers. Besides, about 20% of single-person households showed lower intake rates compared to multi-person households, about 1 to 3 times a month.[2] As such, the intake of nutrients, diet, and food intake in a single household were disproportionate to and miserable than in a multi-family household.

The employment rate of vocational high school graduates, which had been rising since 2010, has also been on the decline since 2017, but the college entrance rate has been on the rise again as the employment rate of vocational high school graduates has dropped to 34.9% since 2018 due to a steep minimum wage hike and the economic downturn.[3] The average college entrance rate is high, at 71.8% and is expected to continue to rise. Due to the nature of universities that are allowed to go to other regions, 33.8% of college students have single-person households.

As such, most high school graduates in the early stages of adulthood entered college, one-person households with one-third of their own, and, as shown in the results above, the school district's eating habits scored the highest with 29.92±5.68 points, while the school district scored 27.24±5.67 points, lower than the dormitory group's. Therefore, this paper seeks to help with the eating culture and eating habits for many college students who are immature in the early stages of adulthood and cannot afford to take care of themselves among single-person households living alone.[4]

2 Preliminary

2.1 HTML, JS, CSS

The most accessible 'web page' was used, and related studies were conducted to use HTML and CSS to create a GUI of a recipe site for college students and JS to include event functions that work on web pages.

2.2 JQUERY

The Jquery is a fast, small, feature-rich JavaScript library. It also makes HTML documents navigation and manipulation, event processing, animation, and Ajax much simpler with easy API that works across multiple browsers. These features give JavaScript code diversity and scalability. Using Jquery's event processing and animation functions, we implemented recipes and chatbot and conducted research.

2.3 Google MAP API

Google MAP API is one of the Open Application Programming Interface provided by Google that allows users to view satellite images around the world. Through API, Google provided functions that developers can customize and utilize, among them, it analyzed the functions of the Geolocation API (GPS). Information about GPS is contained in the window.navigator object in html. One variable object, position.coords, contains not only latitude, longitude, but also speed, altitude, and accuracy, which are implemented to help developers use it. We conducted a related study because we determined that using the above objects could implement appropriate functions to receive information on the device's location and market.

2.4 D3, C3 Library

D3 libraries were used to visualize the various ingredients available to users better. The D3 library has the hassle of drawing even simple charts, but that has the advantage of being able to draw any form. Conversely, when using graphs, visualization was conducted using a C3 library defined in a simple form.

2.5 ChatBot

The chatbot is software that communicates with humans by text or voice. There are many different types of ChatBot OpenAPI at this time, and I wanted to use one of them, the randbot.io. However, for Landbot.io, limitations arise in implementing multiple selections. To solve this problem, the relevant research was conducted by referring to open sources disclosed by CodePen rather than API.

3 Function

3.1 Design

This section describes how the various functions on a web page are designed. Figure 1 below shows the overall structure of a web page. The seven functions and skills required for those living alone who are not familiar with the food culture were designed. HTML, CSS, and JS were used for the overall GUI and function, and the main page called Home was created to make the site's overall structure, and site propensity is known. I tried to use JQuery to show you a recipe that I wanted to teach for those who didn't know much about cooking. To inform the expiration date of the ingredients used in cooking, we wanted to visualize them through D3 library. We thought we would use the C3 library to visualize more and more information about food. A ChatBot was created and designed through a search to inform users of the desired recipe. Also, to show the contents of a nearby mart, Google API GPS will be used to design a function that will show the nearby mart. Random menu functions have been designed through GIFs that are suffering from decision-making difficulties.

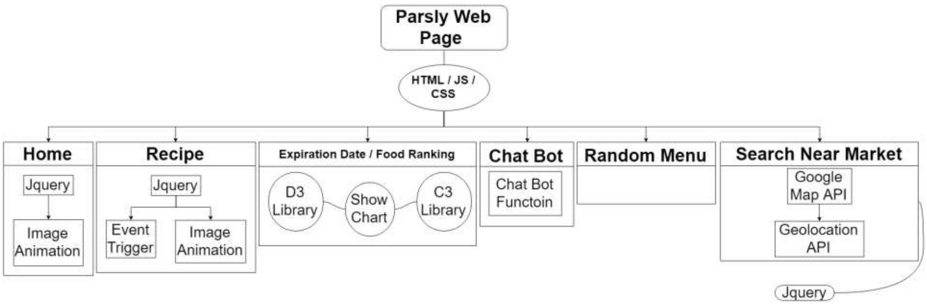


Figure 1. web software architecture

3.2 Implementation

3.2.1 Home

This is the main page you can see when you access the site. The logo and brief introduction are shown centrally so that users can know which site it is. The navigation bar at the top shows various functions of the homepage. The image slide in the center of the screen was implemented through JQuery, and the image was changed after a certain period of time or when a specific button was clicked. It also uses mail to tags on the bottom footer to help facilitate communication between users and managers.

3.2.2 Recipe

We wanted to use jQuery JavaScript libraries to express various images and a lot of text information on a single page. By using event processing and animation provided by jQuery, it not only effectively implemented its functions but also enhanced the readability of recipe descriptions. When switched to the corresponding menu screen, it shows the image, name, brief description, and ingredients of the food. The ID value of the food recipe is assigned to HTML and $(“ID”).show("fast") is executed when the CSS-made button is activated at the bottom of the food image. This performs a call-up of the recipe for which the existing ID value is declared. Print out the recipe on the screen and lower the information in the following menu to the bottom. If the same button is activated again, hide the recipe that was printed on the screen through $(“ID”).hide("fast") and move it to the top.

3.2.3 Expiration Date

To help users store and use food, we want to show information on the shelf life of various ingredients. A D3 library was used because visualizing figures would make it easier and faster for users to understand them. Store the expiration date information of the primary ingredients used primarily in Excel and import the data using d3.csv() in D3. For visualization, Dendrogram, and Grouped Horizontal Bar Chart were used. Use d3.cluster() to create a Dendrogram, a node-link diagram, and use d3.statify() to make each Dendrogram a tree layer. After the ingredients were grouped and divided into individual characteristics, the colors were different for each group to make the visualization more efficient. The number of days of expiration is displayed by creating a horizontal (horizontal) bar graph on the left of the hierarchy. Hovering the mouse over the bar graph also created a circle that showed the exact figure, allowing users to see more accurate figures.

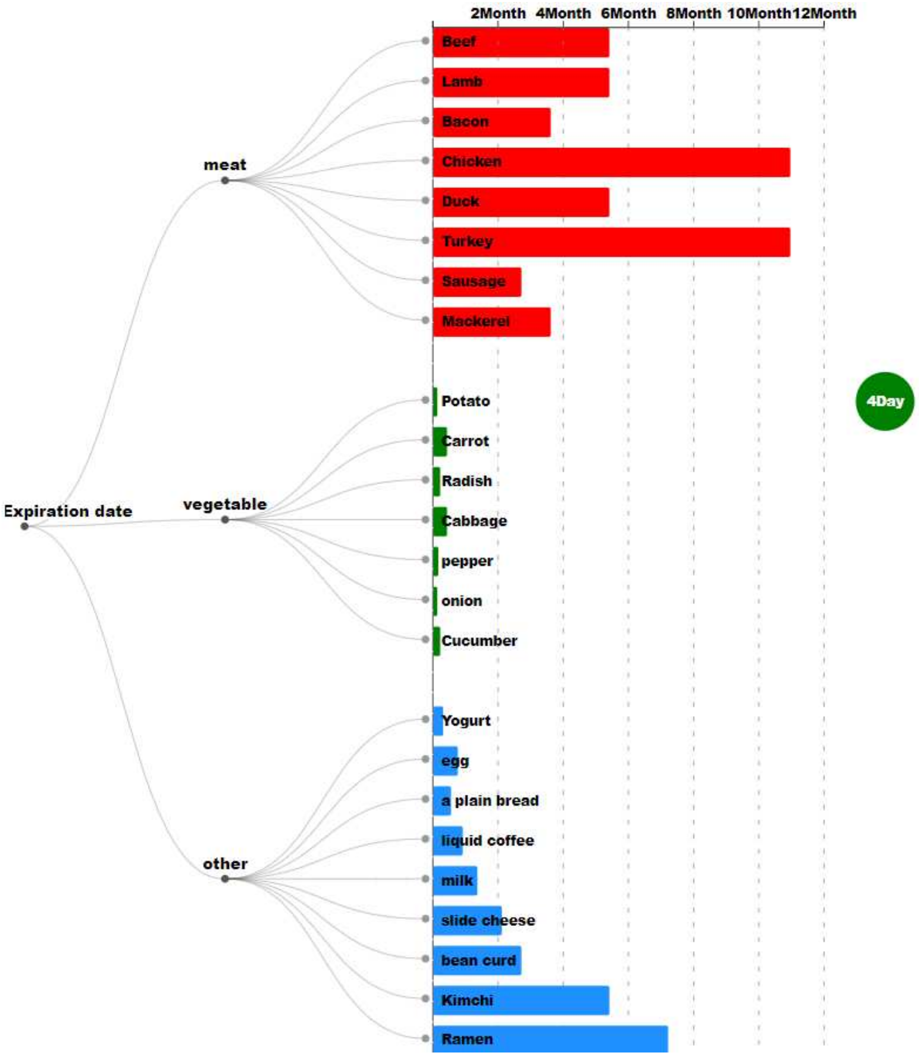


Figure 2. Dendrogram and Grouped Horizontal Bar Chart

This was intended to help users manage and use the ingredients. Needed for the recipe they wanted to make, and then meet the appropriate expiration date.

3.2.4 ChatBot

When recommending recipes, the method was implemented using Chatbot rather than conventional search format.

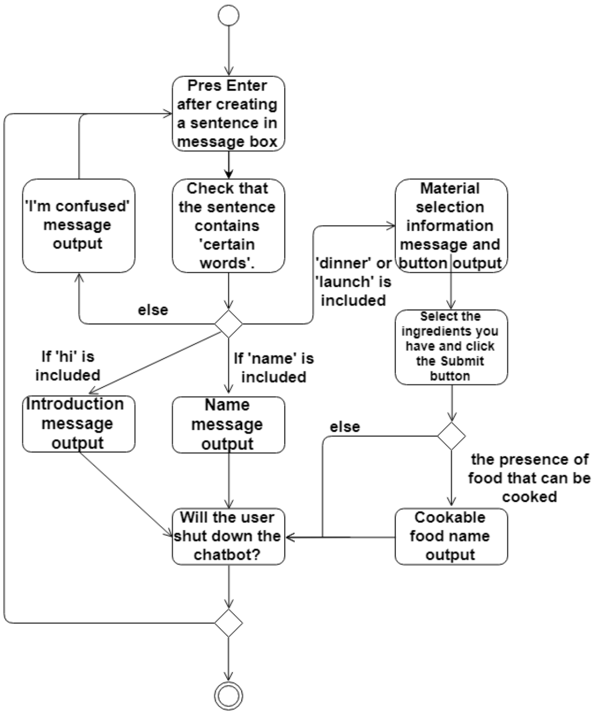
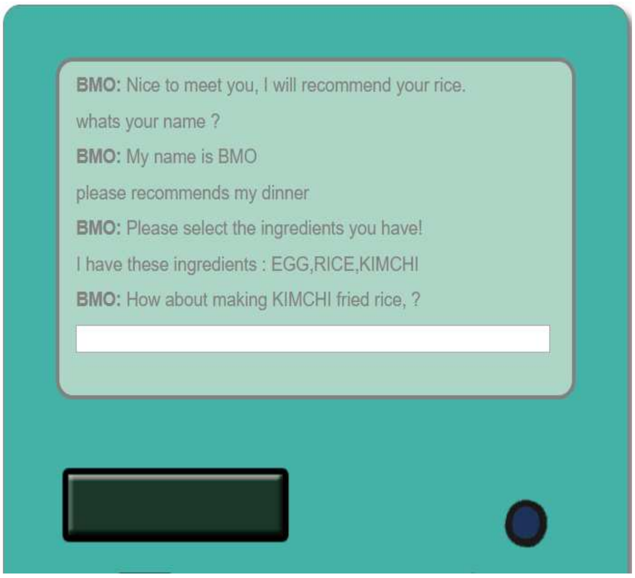

Pres Enter after creating a sentence in message box
'I'm confused' message output
Check that the sentence contains 'certain words'.
Material selection information message and button output
'dinner' or 'launch' is included
else
Select the ingredients you have and click the Submit button
If 'hi' is included
If 'name' is included
Introduction message output
Name message output
else
the presence of food that can be cooked
Will the user shut down the chatbot?
Cookable food name output

Figure 3. Flow chat of Chat bot



BMO: Nice to meet you, I will recommend your rice.

whats your name ?

BMO: My name is BMO

please recommends my dinner

BMO: Please select the ingredients you have!

I have these ingredients : EGG,RICE,KIMCHI

BMO: How about making KIMCHI fried rice, ?

Figure 4. Chatbot Execution Screen

When the chatbot is started, the user enters a sentence in the textbox. The sentence entered in the textbox is tokenized through javascript, and the corresponding answer is printed when the word already defined is included. If you include "dinner" or "lunch," select the ingredients in a button format and print the name of the food you can make through the selected ingredients on the screen. If the word specified for an exception is not included, output "I'm Confused".

The text printed on the screen is raised one sentence by one, and the sentence is deleted if it exceeds the range. A special exit condition upload ends when it leaves the web page.

3.2.5 Searching Near Market

We wanted to receive information from the current device and nearby markets through the Geolocation API (GPS) given in the Google Map API. When a button made of CSS is operated, the [navigator.geolocation.getCurrentPosition] an object within the API receives information about the location. The latitude, longitude, speed, altitude, and accuracy corresponding to the location information are implicit in the [position.coords] object.

Of this information, the latitude and longitude are stored in each variable. If the GPS server is not connected or location information is not received, an error message is printed through exception processing. If you receive location information normally, print the stored latitude and longitude values as a message and print them as an icon on the google map. After that, the map shows the location and information of the existing stored market.

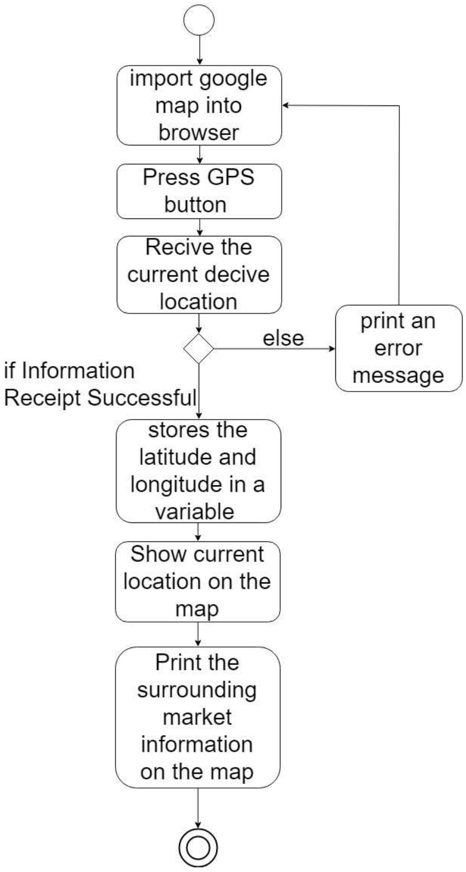

import google map into browser
Press GPS button
Recive the current decive location
else
print an error message
if Information Receipt Successful
stores the latitude and longitude in a variable
Show current location on the map
Print the surrounding market information on the map

Figure 5. Flow chart of GPS function

3.2.6 Food Ranking

To provide users with additional food-related information, we want to show various perspectives on food. For easier and faster user understanding, the C3 library was used to visualize it. C3 is a D3-based library that has made D3's complex charting methods more readily available. We used [c3.generate()] and decided type as a bar, so we visualized the rankings of foods and salty foods that cause the most tooth coloration through a bar graph.

3.2.7 Random

Random Menu was developed to help people who could not easily set the menu. When jQuery was used, the image slider was expressed using the image file of the GIF format, since it was not possible to express the desired amount of image slide speed. This function quickly shows multiple images at 0.1-second intervals when the user clicks the Start button, and when the Stop button is clicked, the user recommends a menu of meals by showing images designated according to the random number generated through javascript.

4. Conclusion

Through the Recipe function of the parsley site, various cooking methods were available for self-taught college students who were not proficient in food culture, and by adding the ChatBot function, users of the Recipe function were more comfortable to use. Through the D3 library, accurate expiration date information was delivered to help prevent health deterioration due to pollution and corruption and additional expenditure due to discarded food ingredients, and through the C3 library, additional information about food was provided to help people understand food culture from various perspectives. They also used the Google Map API to help them buy food ingredients by finding a nearby mart at their location. I could see that extra features that were not provided by other websites were more and more necessary for those who were not used to living alone. Furthermore, we would like to create the various functions that single-person households of all ages need to help more for the growing number of single-person households.

References

1. Yeo Bong Lee, "One-person Households and Their Policy Implications", Health and welfare policy forum, 2017, 252(0), pp.64-77
2. 『Social Security Factbook 2018』. 2019, Ministry of Health and Welfare 사회보장총괄과
3. NaYeon Kang, "Analysis of the Difference in Nutrients Intake, Dietary Behaviors and Food Intake Frequency of Single- and Non Single-Person Households: The Korea National Health and Nutrition Examination Survey (KNHANES), 2014-2016", Korean journal of community nutrition, 2019, 24(1), pp. 1-17
4. Sooram Son (2019. 03. 25), the college entrance rate for 2018 was 69.7%, and the "small rebound" was made. a 'dwindling number' high school graduation rate of 30.7%[news] http://www.veritas-a.com/news/articleView.html?idxno=148146
5. Bokim Lee, “The Relationship of Health Behaviors and Residence Types of University Students”, J Korean Soc School Health, 2012, 25(1), pp. 78-83

Automatic Checking System for Supporting Algorithm Implementation using Open Sources

In Joo, Seung-Woo Kang, Jin-Ho Jeon, Kwan-Hee Yoo
Chungbuk National University, Cheongju, South Korea

Abstract. It is essential to solving algorithm questions by type of subject to learn algorithms systematically. This program provides a useful algorithm learning space for users, such as registering problems, sending messages, step-by-step learning, Re-Solving wrong answered algorithm question and visualizing tendency of type of matter and wrong response rate. This paper introduces the function of the program, the differences from the relevant application, system configuration and flow, web interface, and site operation plan.

Keywords: algorithm, data visualization, compiler, Online Judge

1 Introduction

This program is an Online Judge where users operate autonomously and share information. It is important to learn an algorithm by types of topics to study algorithm systematically. Therefore, various types of algorithm questions and the ability to check the frequency of questions are required. This program provides an effective algorithm learning space for users, such as registering questions, sending messages, step-by-step learning, Re-Solving wrong answered algorithm question, and visualizing tendency of type of question and wrong response rate. This program provides the following function:

· Member system : All members who need to learn algorithms or want to solve questions can utilize the functions of the site through membership and login.

· Question Registration : All members have authority and can be divided into cases where the applicant is an evaluator, and a person is a public person. You can also attach representative images and videos after selecting one type of algorithm. After creating question content, test cases can be inserted based on input and output. This test case can be added several times. The applicant can determine the difficulty of the question and set the rank (easy: 1 ~ difficult: 5).

· Rank System : Rank consists of five in total, and its difficulty varies depending on rank. Each member has a rank and starts at 1. The applicant is not allowed to make questions higher than his rank. A member who solves a question, likewise, cannot solve a question of rank higher than himself. For a member to raise the rank, the member must solve the randomly selected question of each rank to raise the rank to a higher level.

· Algorithm learning : When a member wants to solve an algorithm question, the user can select the desired language from the left side of the page and code it according to the content of the problem. If you're done coding, you can try compiles. On the right-hand side, a random test case will be presented, and the member must fill out the output value based on the input of the test case and submit it. If the answer is correct, the answer message will appear, and if not, the wrong answer message will be displayed, and the question page will appear again.

· A wrong answer note: If any of the questions that a member has solved have been answered incorrectly, it can be gathered and viewed and solved again.

· Message Transfer: This function allows the user to enter the ID of the person to whom the message is sent and send the message. The main purpose is to send and receive messages in the form of e-mail if the member has any questions about the algorithm question to the applicant. And the message is censorable.

· Algorithm Bulletin: This bulletin board allows users to upload and view algorithms and questions freely.

· Visualization of Tendency of Algorithm Question Type : It produces bar charts that have recorded the number of counts for each question type registered by the evaluators.

· Visualization of Wrong Response Rate : The error rate of questions according to the type of algorithm question can be visualized with the scatter plot matrix, and the spot of the matrix can be selected to move on to the corresponding question.

The above functions allow the user to learn the algorithm systematically. The paper proceeds as follows. Section 2 describes the sites involved and their differences. Section 3 describes the program system configuration and flows. Section 4 describes the key features and visualizations Web interfaces. Section 5 concludes and displays future research.

2 Related Study

In this section, we discuss the related sites. There are many Online Judge to learn algorithms.

‘Projecteuler Online Judge’[1] started operation on October 5, 2001. The intended audience includes students for whom the basic curriculum is not feeding their hunger to learn, adults whose background was not primarily mathematics but had an interest in things mathematical, and professionals who want to keep their problem solving and mathematics on the cutting edge. This is a site where users who are interested in math can enter the algorithm with fun.

‘Baekjoon Online Judge’[2] started operation on March 19, 2010. This site can share and study information about different types of algorithms. It also provides many kinds of algorithm questions and various program languages and many other efficient functions.

Besides, there are many other outstanding Online Judge at home and abroad, but the most significant difference from this program is that users can submit questions themselves and study algorithm type by type.

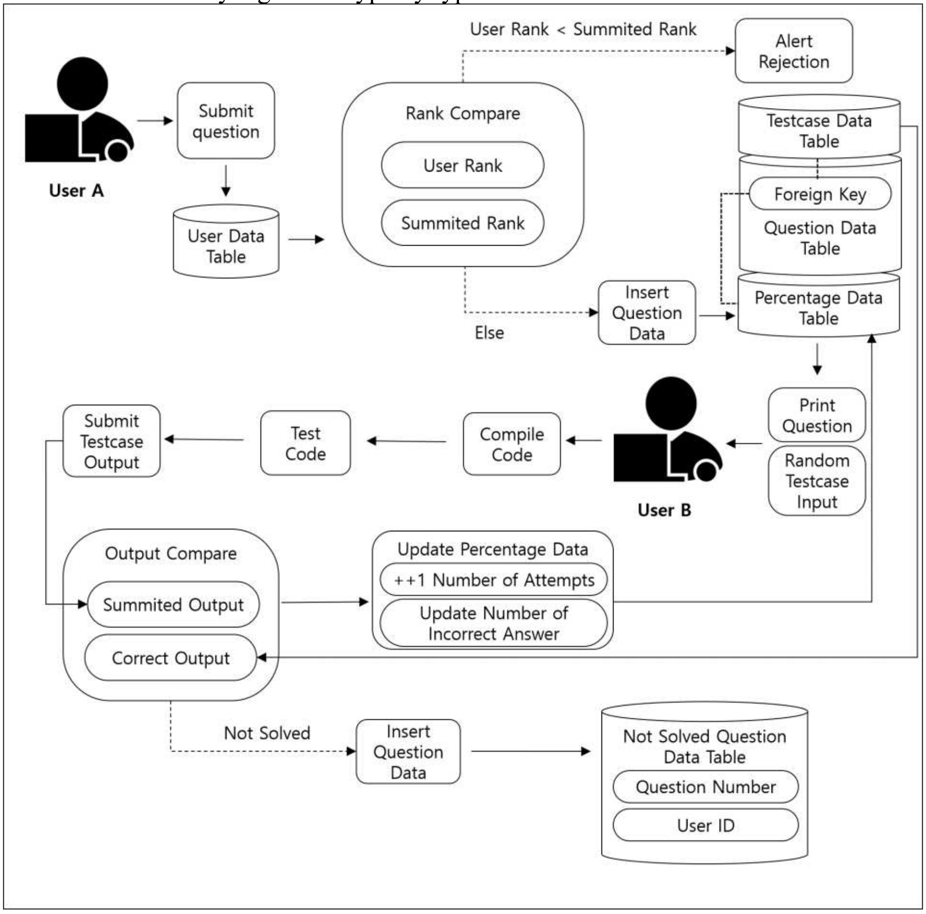


Fig. 1. An algorithm question storing and solving system architecture

3 Program System Configuration and Flow

This section describes the configuration and flow of the program systems that differentiate themselves from existing related programs. Section 3-1 describes the system of storing and solving questions. Section 3-2 describes the system of dealing with wrong answers. Section 3-3 describes user rank management systems through random problem-solving. Section 3-4 describes the system of the visualization system using graphs.

3.1 Algorithm Question Storing and Solving System

In this system, users store or solve problems. If a user finishes writing the question with the test cases, the system loads the rank information from the user data and stores the question if it is not higher than his own rank (if higher, a refusal message will be displayed). Test cases are saved according to the Foreign Key of the question data that is stored. If other users try to solve this question, the system shows the input of random test cases along with the question. The user will test the code if they compile the code according to its content. The user submits the output for the input of the test case, and the system will compare between real output and submitted output of the test case. If matched, this system will be terminated after an update. If not, the system will store the question number and user ID in the 'Not Solved Question' data table. In the end, the system will update the Percentage Data Table, which stores the error rate, depending on whether the answer is correct or not.

3.2 Re-Solving Algorithm Question System

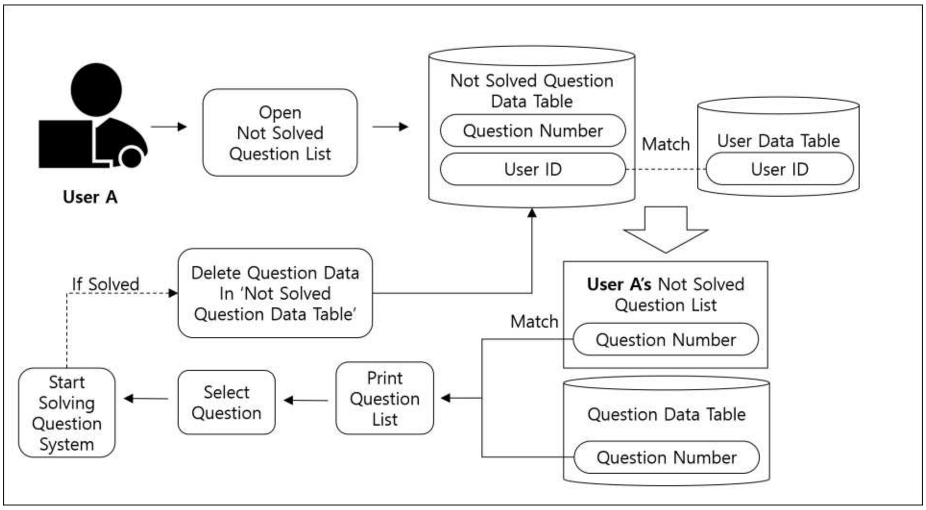

User A
Open Not Solved Question List
Not Solved Question Data Table
Question Number
User ID
Match
User Data Table
User ID
User A's Not Solved Question List
Question Number
Match
Question Data Table
Question Number
If Solved
Delete Question Data In 'Not Solved Question Data Table'
Start Solving Question System
Select Question
Print Question List

Fig. 2. A Re-Solving algorithm question system architecture

The system allows users to print out, select, and solve wrong questions. The system will create a list of question number that matches the user ID of the Not Solved Question data table and the user ID of the user data table. The system will compare the list with the question number of the question data table and print the question list to the user. Within the printed list, users can select and solve problems. The system configuration and flow to solve the problem are the same as section 3-1, and if correct, delete the data from the Not Solved Question data table.

3.3 Ranking System

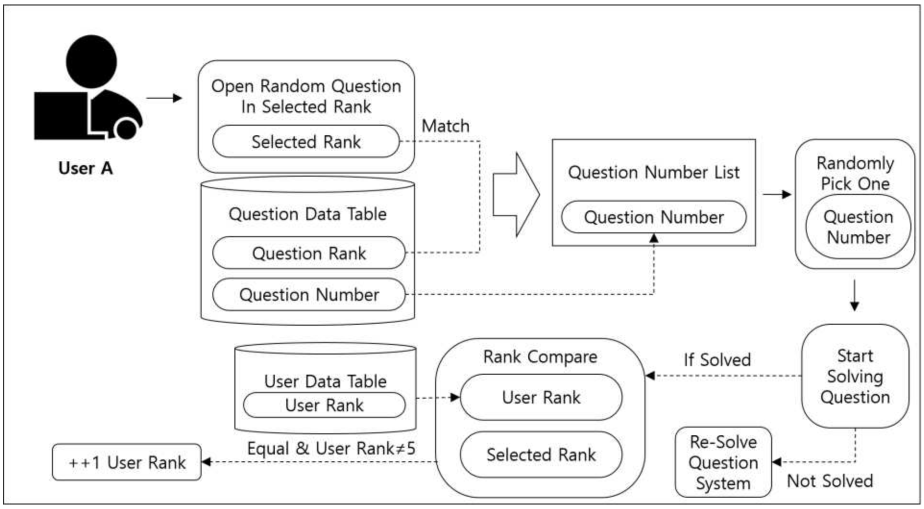


Fig. 3. A ranking system architecture

This system raises the user's rank by allowing users to solve random questions. The system compares the ranks selected by the user with those in the question data table and creates a list of question numbers with those matching. Among them, a random question is chosen so that users can solve it. The process of solving a question is the same as the Section 3-1 system. If it is not solved, it is handled as in Section 3-1. If the submitted answer is correct, the rank is raised one notch higher and stored in the user data table when the following two conditions are met:

· The user's rank in the user data table is the same as that of the user-selected rank.

· The user's rank in the user data table is not five which is maximum.

3.4 Visualizing a Tendency of Question Type & Wrong Response Rate System

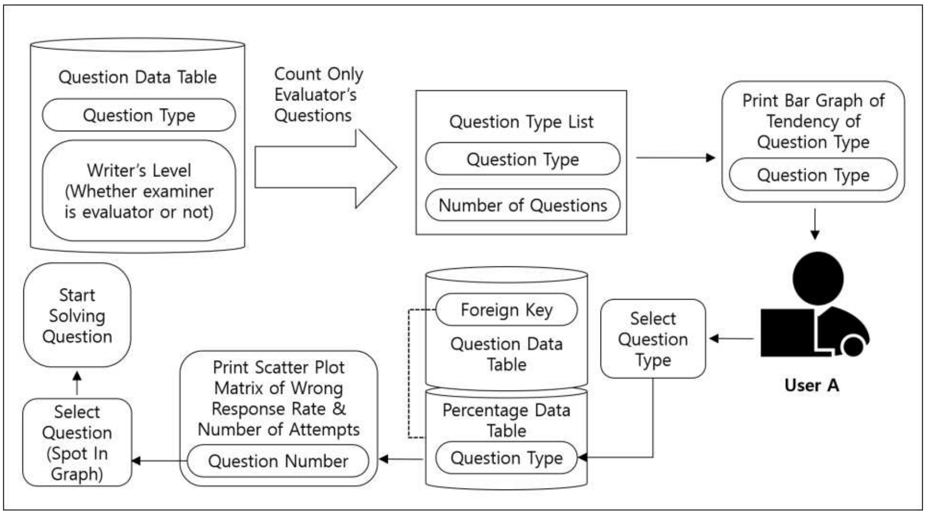
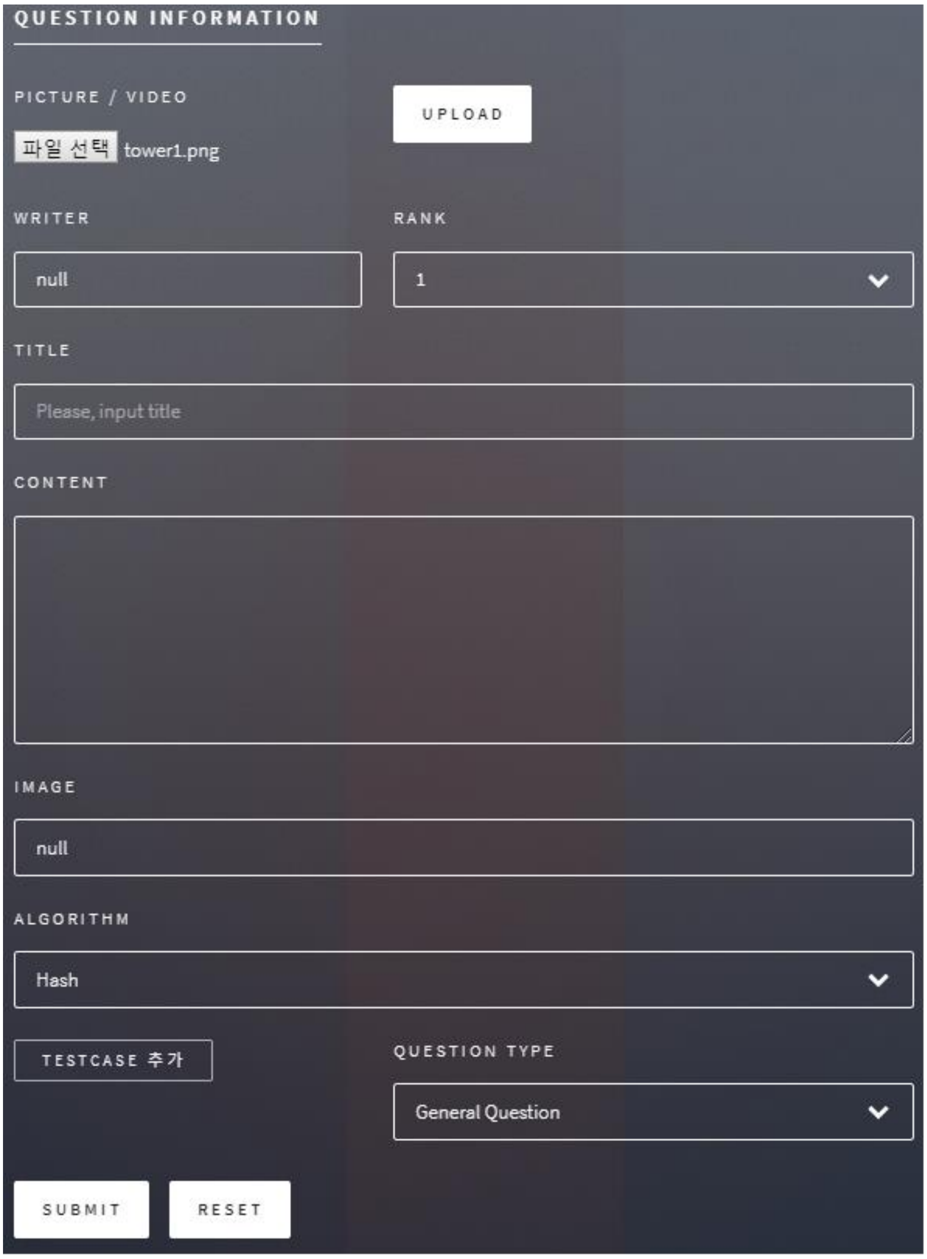

Question Data Table
Question Type
Writer's Level (Whether examiner is evaluator or not)
Count Only Evaluator's Questions
Question Type List
Question Type
Number of Questions
Print Bar Graph of Tendency of Question Type
Question Type
User A
Select Question Type
Foreign Key
Question Data Table
Percentage Data Table
Question Type
Print Scatter Plot Matrix of Wrong Response Rate & Number of Attempts
Question Number
Select Question (Spot In Graph)
Start Solving Question

Fig. 4. A visualizing a tendency of question type & wrong response rate system architecture

The system graphically visualizes the trends in questions from the evaluator and the error rate of questions. The system counts question data in the question data table, which examiner level is an evaluator. Then, it creates a list of question type which contains the number of question. Finally, using C3, the system creates a bar graph and show it to the user. When a user selects the problem type, the system takes the error rate, and the number of attempts from the percentage data table and visualizes it by turning it into a scatter plot matrix. When a user hovers over each point of the matrix, the system access the question data table using the foreign key and print out the problem information. When clicked, solving the question system starts. The process of solving a question system is the same as section 3-1.

4 Main Function & Web Interface

In this section, we discuss the main functions and the Web interface.



QUESTION INFORMATION

PICTURE / VIDEO

tower1.png

WRITER

null

RANK

1

TITLE

Please, input title

CONTENT

IMAGE

null

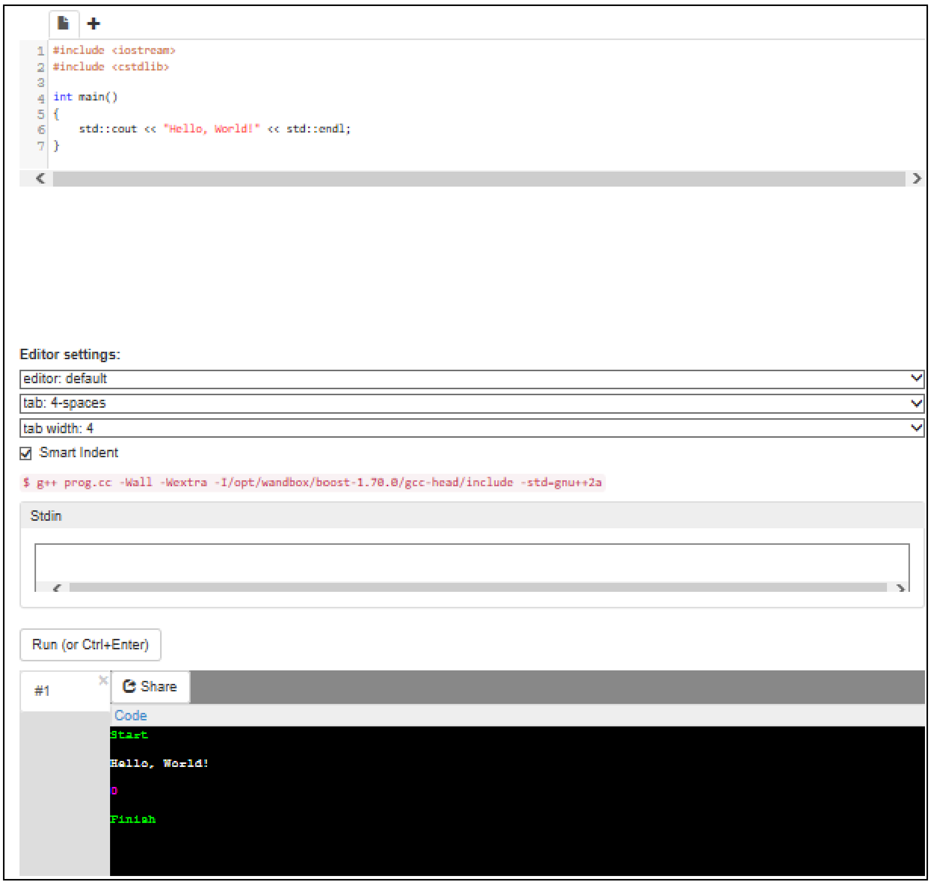
ALGORITHM

Hash

QUESTION TYPE

General Question

Fig. 5. An interface of question registration



```
#include <iostream>
#include <cstdlib>

int main()
{
    std::cout << "Hello, World!" << std::endl;
}
```

Editor settings:

editor: default

tab: 4-spaces

tab width: 4

☑ Smart Indent

```
$ g++ prog.cc -Wall -Wextra -I/opt/wandbox/boost-1.78.0/gcc-head/include -std=gnu++2a
```

```
Start

Hello, World!

0

Finish
```

Fig. 6. An interface of question solving (left-hand side of page)

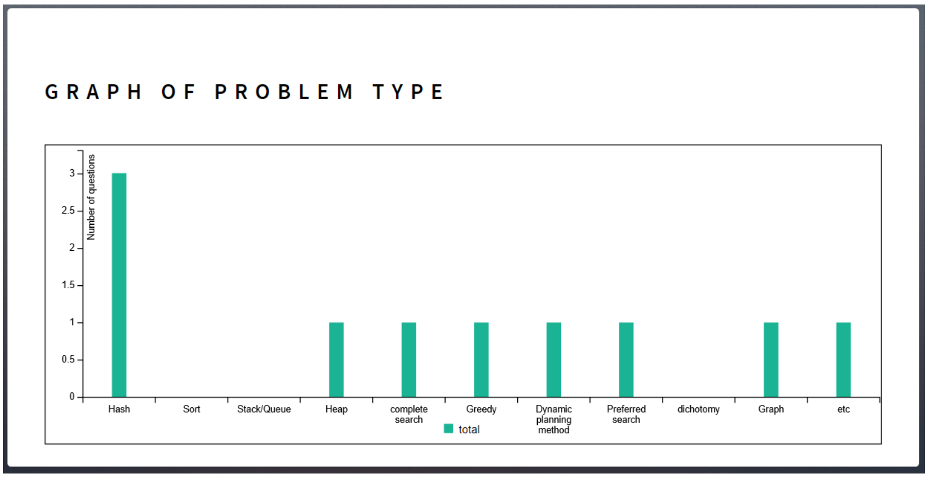


Fig. 7. An interface of visualization of a tendency of question type

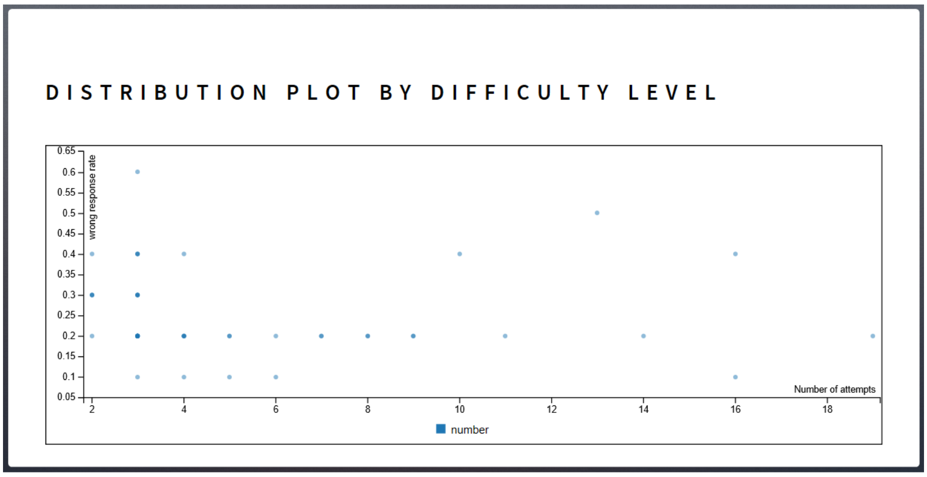


Fig. 8. An interface of visualization of wrong response rate

Figure 5 shows that users can register an algorithm question. The users can register their algorithm problems freely and conveniently through the interface provided. Figure 6 shows that users can solve a registered question. The users can use compilers to solve algorithm question and register output as an answer of the question. Figure 7, 8 shows a graph of data visualization. The users can see a tendency of question type and wrong response rate in a bar graph and scatter plot matrix.

5 Conclusion

Learning algorithms by type and empowering all users to register problems is critical to systematically learning algorithms and freely sharing algorithm knowledge. This program makes it possible to do them. Also, it provides a Re-Solve system that enables effective and repetitive learning of algorithms. Moreover, this program makes it easy for users to learn algorithms with a convenient and simple interface. Therefore, this program provides a more systematic and convenient algorithm learning space. Future maintenance plans include:

- board management
- compiler Improving
- members date analysis and type recommendation

References

1. [1] Project Euler, "About", 2019.07.08, https://projecteuler.net/
2. [2] Baekjoon Online Judge, "About", 2019.07.08, https://www.acmicpc.net/about

Customizable Timetable Generator Using Web Crawler

JiHyun Na, Jihyun.dev@gmail.com
MoonHyun Kim, moonhyun.dev@gmail.com
KyeongJun Kim, person474845@gmail.com
In Joo, jooin95@naver.com
Kwan-Hee Yoo, khyoo@chungbuk.ac.kr
Department of Computer Science, Chungbuk National University, South Korea

Abstract. There are so many sites or application to help user customize their timetable. Most of these tools are working well. But the problem is whether the data they have and real information are the same. There are many ways to get information to do that, but it is most accurate to get it directly from an institution that has information. In the end, information is obtained through students attending the school because not all institution is happy to give it to them. In this process, data inaccuracy occurs. When it comes to scheduling timetable, the accuracy of the data is important. So in this paper, we look at method how to get the information without data inaccuracy. We used a web crawler to store new data or update data. Through this, we can provide exact data to the user and help them customize their timetable according to their needs. Other than that, we provide such as community function among members, visualization timetable. In this paper, we introduce the function of the program, differences from existing programs, flow chart of program, web-interface. Before entering the text, the first thing we want to say is that it's not a solution. Because we can't access the anther school system, this work is focused on particular school systems. Ultimately, we point out possible approaches for similar cases.

Keywords: web crawler, timetable, data visualization

1 Introduction

This work began from two perspectives on the platform that often used today. One is data inaccuracy, and the other is a waste of time and effort. Whether you are a student now or have already graduated, you may have made a timetable for once. Making a timetable that suits each one is important for a good semester scheduling, but it can be bothersome because it takes a lot of time and effort to make efficient timetable considering various user requirements. There are many site or application to help to make a timetable. When you choose the platform to make a timetable, then you've probably seen that it's different from the actual information. As we said in Abstract, because most of them can't get information from the institution directly, this problem occurs. A single change in information can change the overall timetable. A timetable created through time and effort can become obsolete. The second perspective begins here.

There are dozen of choices to make one timetable. You should also look up the timetable for the various classes and consider whether they can be included in my timetable. And you have to make another timetable because you may not be able to apply for classes according to the time table you created first. Even if you've completed your desired timetable, if the time changes in class you selected, your time and effort will be wasted.

So in this work, we considered how to maintain the accuracy in data on our platform and how to minimize time and effort to make a timetable. This work includes the following functions:

· Data crawling: Through the crawler, this platform can have exact data and serve the services to the user.

· Customize the class requirement: User can receive the result filtered by customized information of their choices such as major, grade, and desired number of credits according to their needs.

· Customize the user requirement: In addition to class requirement, the user can set a priority of lecture or can block the time in scheduling according to their needs.

· Auto Generate Timetable: User can make timetables as they like without much time and effort.

· Result visualization: This program provides visualized timetable created by result schedule and show this on web-interface.

· Community & Membership: Every user who wants to share information related with lecture can utilize community through the join in membership and logging in.

· Store result: User can access and update the visualized result timetable by storing in their information section.

This report is organized as follows.
Section 2 looks at existing approaches in another platform.
Section 3 describes how-to approaches to solving existing problems
Section 4 shows how this program works with the flowchart
Section 5 shows how the interface is structured
Section 6 concludes.

2 Related Study

There are already many web sites and application that users manually make a timetable and show the result by visualizing. Among them, one of the representative sites is ‘Every Time’ that have been providing the services sines 2011.

This site has over 3.6millon members from over 400 universities and colleges. In the beginning, users in this site had to register lecture information about the university in which they belong to get a visualized timetable. After that, they were able to collect the data using Open API provided only for some universities. So, they can offer the lecture information to the user. However, in the case of most of the school, the web manager can still collect information manually from each university user.

So they get information from students attending a school and register it their sites or application. They ask to them give them the class information. Or they wait until the user upload the information on their platform and after that, they share to the user. If it works this way, it doesn’t seem to be a problem, but actually, in the class registration period, so many changes occur for many reasons. To provide more accurate and better services, they have to check whether there is a change or not in class information. If there is a change, then they ask the user to give them the information again, or there is no choice but just to wait until the user revises the information. But this period is only a week or so, and changes also occur frequently during that short period. So most site and application couldn’t import these data on time. Because of this reason, many users confused because of data difference between data on sites or application and original sites data operated by their school.

So we used Crawling to solve this problem. More information about that is described in Section 4 below.

3 Method

This work automatically creates a timetable for various requirements to save time and effort. For this, it should be able to import a large amount of data accurately and easily. This program collects data related to the major, a lecture from the webserver using web crawler and then, using minimize error caused by user when they make their timetable. We used Java Selenium package and Chrome Driver for Crawling. Selenium is a portable framework for testing web applications. Through this, we can access the web site has the information we want. It is working on the Chrome browser, and it looks for a tag on the web page sources, and it gets that information or inputs data on that tagged place. Example, if you need to input ID and Password then write the information and tag name you want to give in your code. Then the information is delivered during runtime if the code is working well. In this way, the program can access where the data is stored.

In this program, it searches data by department and extract data separately by the department and store it as an Excel file. After that, Using SQL statement, it stores Excel file data in the database. When using crawling, Traffic problems can occur in this process. To avoid this problem, automatically log in every hour to see if there are any changes in information. This way, data can be kept up to date

A detailed description of each step is described in the following Section 5, along with the Flow chart.

4 Program Flowchart

In this section, we explain the process of data collecting and data manipulation. Figure 1 demonstrates overall program flow, and Figure 2. demonstrates the customizing process by user requirements.

As shown in Figure 1, At first, crawler store data brought from the web server in a local computer. At this time, data from the webserver is stored in local in the form of excel files. Each excels files have information like lectures of one major or electives of one field. Stored excel file's lecture information is stored to one schema in the local database through the store process. After that, On the web-interface, the program receives input from the user. And based on these requirements, the result by the customizing process is shown up to the user.

- **4.1 System Flowchart**

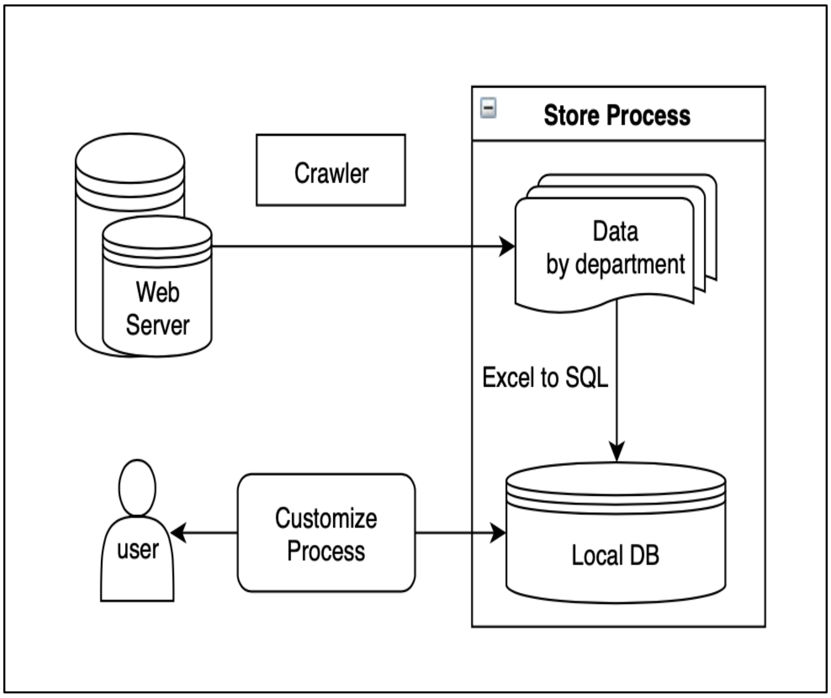


Figure 1. System Flowchart

- **4.2 Customizing Process**

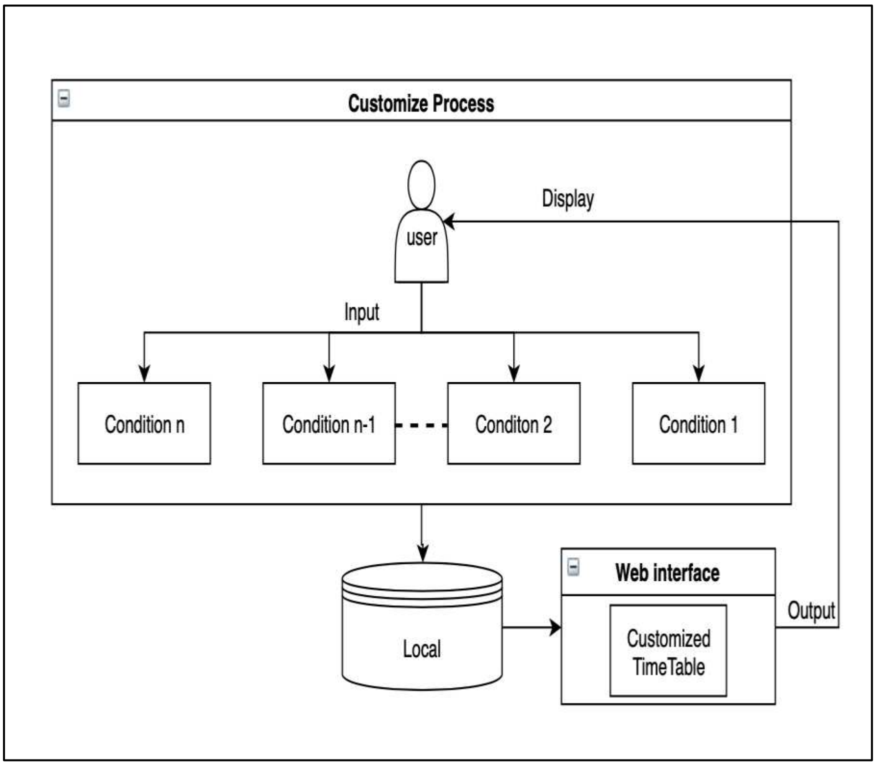


Figure 2. Customizing Process

Figure 2. demonstrates the process that the data received from user is how it's handled. At first, user search for lectures they want to take in. After that, user select the lecture and add to their timetable. At the same time, users can set block time or some constraints in their timetable if they don't want to take a class in particular time or day. For example, before making a timetable, users have to enter information about the classes they have already taken. That way, they can prevent courses that you have already received from being added to the timetable. If they select their major, then they will see a list of existing classes in the current semester. And then the users choose the least of the classes they want to take. At the same time, if there is a day they don't want have a class, then they can block the day in order to avoid adding classes on that day of the week.

After these steps, the system shows the timetable considered by lectures user already took in or prerequisite along with existing constraints. Space, where major classes were not filled, will be added with elective class information they can take in. Users can store this result on their personals section, so access and update are available whenever they want.

5 Web Interface

In this section, we describe the prototype of a Web interface that visualizes the customized timetable by user's conditions. Each figure shows the process of creating the result.

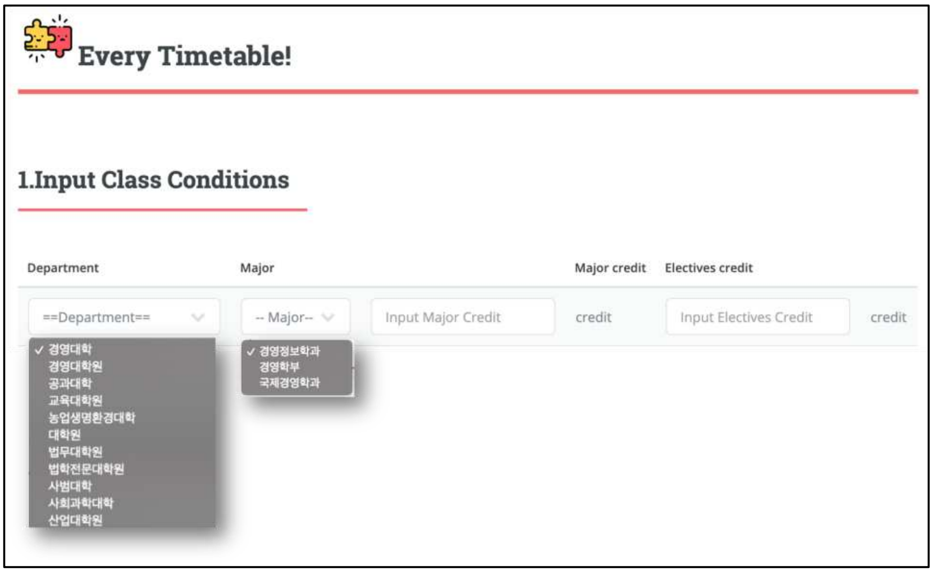


Figure 3. Input form class conditions

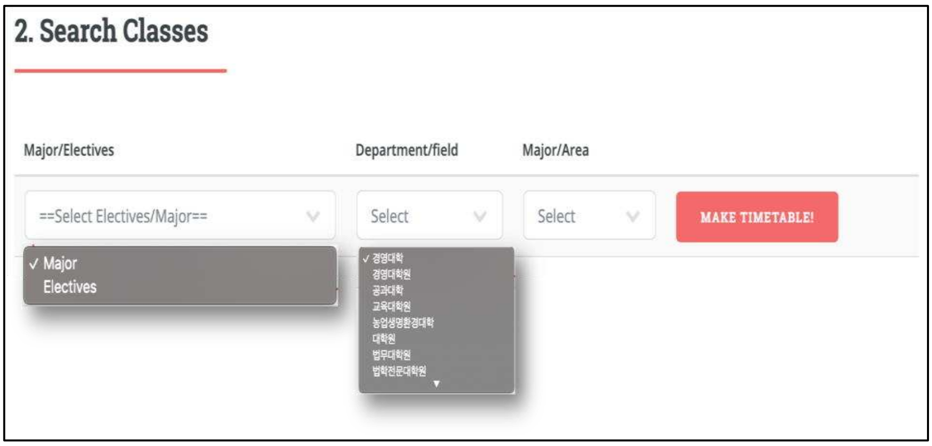
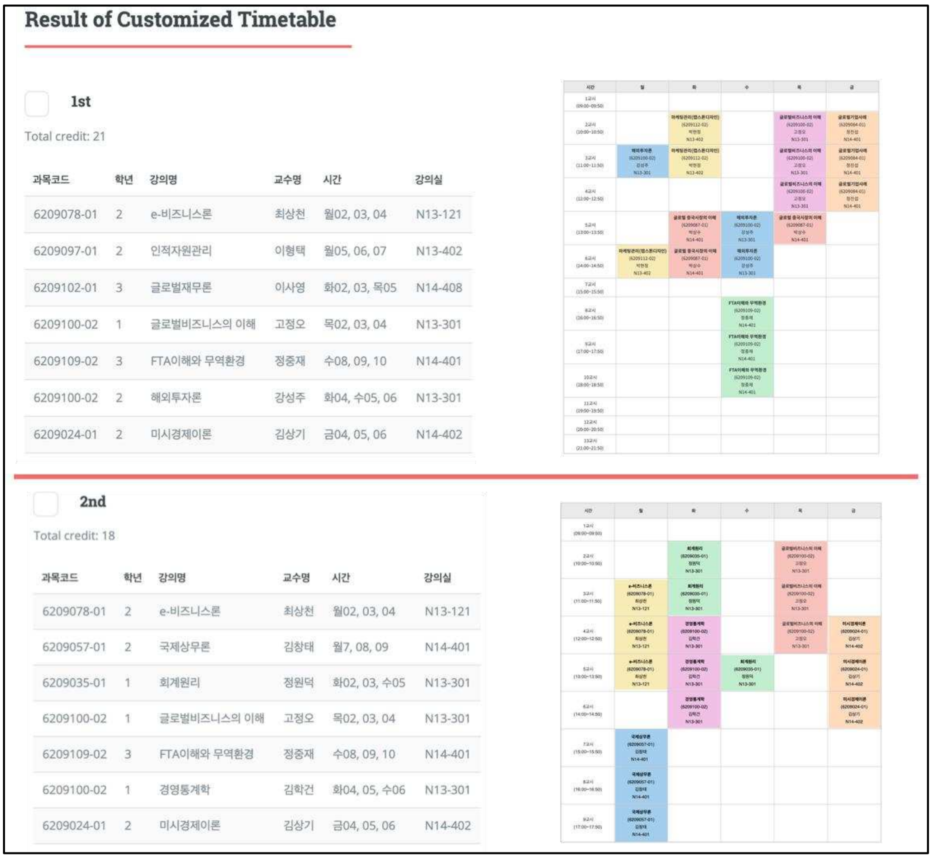


Figure 4. Search classes



Result of Customized Timetable

□ 1st

Total credit: 21

과목코드	학년	강의명	교수명	시간	강의실
6209078-01	2	e-비즈니스론	최상천	월02, 03, 04	N13-121
6209097-01	2	인적자원관리	이형택	월05, 06, 07	N13-402
6209102-01	3	글로벌재무론	이사영	화02, 03, 목05	N14-408
6209100-02	1	글로벌비즈니스의 이해	고정오	목02, 03, 04	N13-301
6209109-02	3	FTA이해와 무역환경	정중재	수08, 09, 10	N14-401
6209100-02	2	해외투자론	강성주	화04, 수05, 06	N13-301
6209024-01	2	미시경제이론	김상기	금04, 05, 06	N14-402

시간	월	화	수	목	금
1교시 (09:00~09:50)					
2교시 (10:00~10:50)		마케팅관리(캡스톤디자인) (6209112-02) 박현정 N13-402		글로벌비즈니스의 이해 (6209100-02) 고정오 N13-301	글로벌기업사례 (6209084-01) 정진섭 N14-401
3교시 (11:00~11:50)	해외투자론 (6209100-02) 강성주 N13-301	마케팅관리(캡스톤디자인) (6209112-02) 박현정 N13-402		글로벌비즈니스의 이해 (6209100-02) 고정오 N13-301	글로벌기업사례 (6209084-01) 정진섭 N14-401
4교시 (12:00~12:50)				글로벌비즈니스의 이해 (6209100-02) 고정오 N13-301	글로벌기업사례 (6209084-01) 정진섭 N14-401
5교시 (13:00~13:50)		글로벌 중국시장의 이해 (6209087-01) 박상수 N14-401	해외투자론 (6209100-02) 강성주 N13-301	글로벌 중국시장의 이해 (6209087-01) 박상수 N14-401	
6교시 (14:00~14:50)	마케팅관리(캡스톤디자인) (6209112-02) 박현정 N13-402	글로벌 중국시장의 이해 (6209087-01) 박상수 N14-401	해외투자론 (6209100-02) 강성주 N13-301		
7교시 (15:00~15:50)					
8교시 (16:00~16:50)			FTA이해와 무역환경 (6209109-02) 정중재 N14-401		
9교시 (17:00~17:50)			FTA이해와 무역환경 (6209109-02) 정중재 N14-401		
10교시 (18:00~18:50)			FTA이해와 무역환경 (6209109-02) 정중재 N14-401		
11교시 (19:00~19:50)					
12교시 (20:00~20:50)					
13교시 (21:00~21:50)					

□ 2nd

Total credit: 18

과목코드	학년	강의명	교수명	시간	강의실
6209078-01	2	e-비즈니스론	최상천	월02, 03, 04	N13-121
6209057-01	2	국제상무론	김창태	월7, 08, 09	N14-401
6209035-01	1	회계원리	정원덕	화02, 03, 수05	N13-301
6209100-02	1	글로벌비즈니스의 이해	고정오	목02, 03, 04	N13-301
6209109-02	3	FTA이해와 무역환경	정중재	수08, 09, 10	N14-401
6209100-02	1	경영통계학	김학건	화04, 05, 수06	N13-301
6209024-01	2	미시경제이론	김상기	금04, 05, 06	N14-402

시간	월	화	수	목	금
1교시 (09:00~09:50)					
2교시 (10:00~10:50)		회계원리 (6209035-01) 정원덕 N13-301		글로벌비즈니스의 이해 (6209100-02) 고정오 N13-301	
3교시 (11:00~11:50)	e-비즈니스론 (6209078-01) 최상천 N13-121	회계원리 (6209035-01) 정원덕 N13-301		글로벌비즈니스의 이해 (6209100-02) 고정오 N13-301	
4교시 (12:00~12:50)	e-비즈니스론 (6209078-01) 최상천 N13-121	경영통계학 (6209100-02) 김학건 N13-301		글로벌비즈니스의 이해 (6209100-02) 고정오 N13-301	미시경제이론 (6209024-01) 김상기 N14-402
5교시 (13:00~13:50)	e-비즈니스론 (6209078-01) 최상천 N13-121	경영통계학 (6209100-02) 김학건 N13-301	회계원리 (6209035-01) 정원덕 N13-301		미시경제이론 (6209024-01) 김상기 N14-402
6교시 (14:00~14:50)		경영통계학 (6209100-02) 김학건 N13-301			미시경제이론 (6209024-01) 김상기 N14-402
7교시 (15:00~15:50)	국제상무론 (6209057-01) 김창태 N14-401				
8교시 (16:00~16:50)	국제상무론 (6209057-01) 김창태 N14-401				
9교시 (17:00~17:50)	국제상무론 (6209057-01) 김창태 N14-401				

Figure 5. Customized timetable result

Figure 3. shows input forms of search conditions for classes. The users can customize their conditions through these input forms. In this interface, First, the users can set the number of credits that should be included in the timetable. Then next, the users can set class conditions that consist of the department, major, classification, grade, etc. Each of the conditions is stored in the database by the web crawler.

If you select a specific condition, you will see the list of results that are searched by conditions. You can also add a specific class by searching the class using the search form in Figure 4. After you select each input, the list of classes is loaded from the database. You can select classes that you want to add to your timetable using a checkbox.

If the user selects classes and clicks the 'make timetable' button, then the users can see their customized timetable as shown in Figure5. Figure 5. shows customized timetable of various cases. Also, those are shown by the priority that the user selected conditions (Figure 3, 4). This result depends on the user's conditions. To easily distinguish, each of the classes has a randomly colored background. Each block shows class name, id, professor, room number. If the user chooses a duplicated condition, through the customizing process (Figure 2.), the classes are not shown on the timetable.

6 Conclusion

It is a hassle to make timetable considering all of your conditions. Not only it takes a lot of time and effort, but also it is difficult to get accurate and latest class data. So, in this paper, to keep the data up-to-date, we collect the data from the web using a crawler. Besides, if we can automatically customize various conditions, this will save your time and effort. Furthermore, this web site provides a user with the visualized timetable and community to share their timetable. Overall, this website will make your time more efficient and convenient.

Reference

1. [1] Every Time, "About", 2019.07.15, https://everytime.kr/

Development of a cosmetics component analysis app that enables real-time image recognition

Da-bin Choi, Mi-seon Kim, YangMingFei, , Sang-hyun Choi,

Dept. Management Information System, Chungbuk National University,
Cheongju, South Korea
{choidb1018, sosgik96,mingfei84}@naver.com,{ chois}@cbnu.ac.kr

Abstract. In this paper, the idea of app development, which provides consumers with a level of cosmetic components analysis based on text extracted through real-time image recognition, was written. When users want to know the safety level of cosmetic products in the store, they can quickly and briefly obtain the information what they want by recognizing the images real-time with a smartphone camera.

Keywords : Cosmetics Components Analysis, Cosmetic Safety, Real-Time Image Recognition, Text Extraction

1 Introduction

With controversy over the components of cosmetics, consumers are demanding information on the safety of cosmetics and objective evaluation. As a result, demand for app services that can obtain component analysis information on cosmetics is increasing. For apps on the market, it is inconvenient for users to know the exact name of the product and enter it in person.

Therefore, we recognized the need for app to simplify the search function of the product and to intuitively show safety information about the product. In this paper, the image processing deep learning technology and OCR technology were applied to the cosmetic component analysis service. So, an app service was designed to let users know the safety level of a product by searching for images.

2 Related Research

A similar app service called ‘Hwa-Hae’ is a free app provided by ‘Bird View’ in the cosmetics category of Google and Apple's App Store.

This app provides analysis information on components of cosmetics, but it has problems such as inconvenience of search function that requires direct input of brand and product name. Search capabilities, which are the basic requirements of the information-delivery app, have been implemented without consideration of user diversity, so it has limitations on service accessibility.

3 The main function (of app)

The app pre-processing extracts text areas from images in real-time through OpenCV techniques when users recognize cosmetics through their camera and converts the images into text through OCR processing by Google Cloud Platform.[1] Visualize the final safety level and the specific components level of each product by comparing the converted text with the product information database. Use the EWG level list and components dictionary established in KCA for components level. This allows users to search safety information about the product quickly without entering the product name.

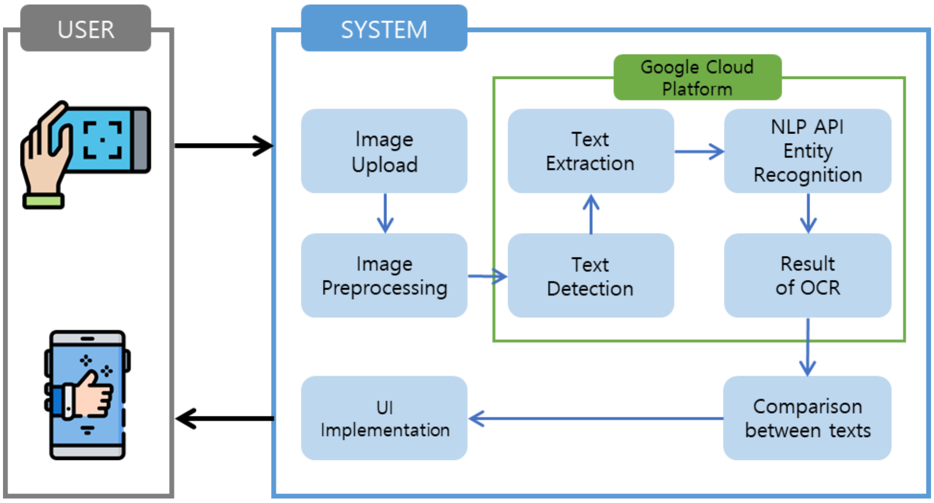


Fig. 1. Our App’s Data Flow

4 Conclusion

The development of smartphones used by most people has led to the development of the application industry. At this point, Apps from various fields are under development, and among them, consumers will demand and choose one that is more convenient to use.

In this paper, the app was designed to be searchable by applying the real-time image recognition function by differentiating it from the existing similar app that requires accurate product name entry. Based on the aforementioned functions, convenient search functions are provided, allowing users to more intuitively know information such as the safety level and specific components of cosmetics with a single recognition of the product.

Reference

1. Sung-eun Chae, Ki-seok Jung, Jeong-yeol Lee, Young-J. Rho. (2018). Development of Smart Household Ledger based on OCR. The Journal of the Institute of Internet, Broadcasting and Communication, 18(6), 269-276.

2017 National Convergence R&D Data Analysis

Tae-hee Lee[1], Jang-won Lee[2], Ji-ho Choi, Jun-hwan Lee[2],

[1] Department of Bigdata, Chungbuk National University,
Cheongju, South Korea
[2] Department of Manangemnt Information System, Chungbuk National University,
Cheongju, South Korea
{taehee8264}@hanmail.net, {wkddnjs416, okcjhh}@naver.com, {okcjhh}@naver.com;

Abstract. With the advent of the Fourth Industrial Revolution, Convergence of various technologies is being made to create new added-value. As data-driven R&D becomes more active, Data analysis is being highlighted as a major driving force for research and development. In this study, we look at domestic R&D convergence industry research trends. And Samples of national R&D projects conducted in 2017 will be analyzed based on 61,280 data. First, we will do a basic analysis of the entire data and compare the single task and the convergence task to the R&D according to the task classification. Finally we will look at the analysis focusing on performance which is a paper and patent

Keywords: Data Visualization, National R&D, Convergence R&D Introduction

1.1 Purpose and Background of Research

Recently, Interest in the fourth industry is very high. The press is reporting on the Fourth Industrial Revolution and also the government is preparing policies to prepare for this. According to Klaus Schwab's Fourth Industrial Revolution, the fourth industrial revolution is the convergence of cutting-edge technologies such as AI. The key element of those technologies is the convergence of various individually developed technologies. The technologies which is Digital, bio and offline-technology converge into various new forms to create new added value. Another major feature is ‘speed’. When new things or techniques are invented or discovered, the pace of its spread is incomparably faster than in the past.

We notice that this fourth industrial revolution can have a huge impact on all areas. According to the Ministry of Science Technology Information Communication, all over the world, enabled data-driven R&D is emerging as a major driving force for data analytics R&D. This phenomenon stimulated convergence and joint research in multiple fields. So, the need for research data sharing and utilization systems is increasing. Major countries are taking note of these changes and implementing an open science policy that opens up research results and processes. However, related systems and infrastructure are insufficient, such as not managing research data in South Korea as a result. In response, Ministry of Science Technology Information Communication is working on a policy plan with the aim of establishing a research data sharing and utilization system.

2 Research Trends of Converged R&D Industry in Korea

2.1. A Study on the Strength of SW Converged R&D by Delphi Analysis

30 experts within the industry, academic, and annual on SW R&D judged the SW-converged R&D strength of the task.
First of all, we decided on the criteria for SW convergence R&D. We divided them into five distinct levels: High, Med High, Medium, Med Low, and Low. Each distinct level was determined based on task title, summary information, and task performance data. And the importance of SW-converged R&D activities was determined for each task.

가. Overall SW Convergence R&D Status

(Unit: %, One million won)

Sortation	**sampling sample**	**consensus scale**				total
		High	Med High	Medium	Med Low	
Total research cost (considering strength)	750,965	18,158	16,576	20,776	9,594	65,104
the specific		27.9%	25.5%	31.9%	14.7%	8.7%

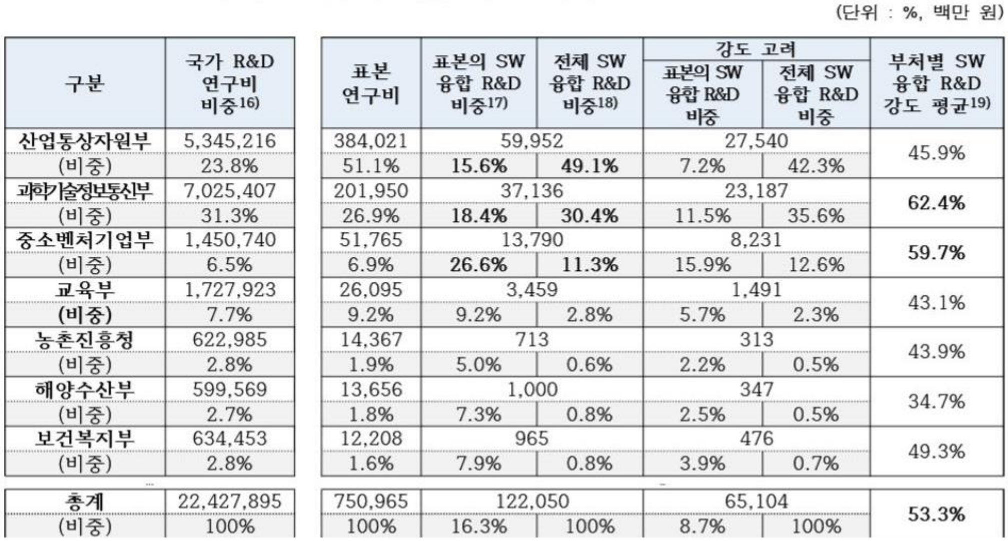
gravity						

- [table 1] Number of SW Converged R&D Tasks and Results of Research Cost

Of the 750,965 million won extracted, the amount corresponding to the scale given is as shown in [Table 1]. The above data is intended to be used to estimate the research costs spent purely on SW convergence R&D.

나. SW Convergence R&D Characteristic Analysis by Type

1) An administrative department



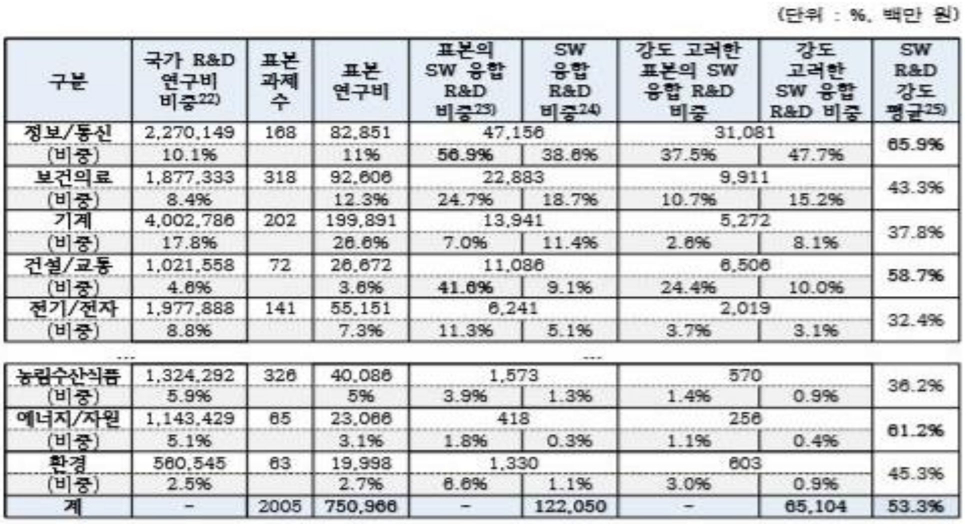
(단위 : %, 백만 원)

구분	국가 R&D 연구비 비중16)	표본 연구비	표본의 SW 융합 R&D 비중17)	전체 SW 융합 R&D 비중18)	강도 고려		부처별 SW 융합 R&D 강도 평균19)
					표본의 SW 융합 R&D 비중	전체 SW 융합 R&D 비중	
산업통상자원부	5,345,216	384,021	59,952		27,540		45.9%
(비중)	23.8%	51.1%	**15.6%**	**49.1%**	7.2%	42.3%	
과학기술정보통신부	7,025,407	201,950	37,136		23,187		**62.4%**
(비중)	31.3%	26.9%	**18.4%**	**30.4%**	11.5%	35.6%	
중소벤처기업부	1,450,740	51,765	13,790		8,231		**59.7%**
(비중)	6.5%	6.9%	**26.6%**	**11.3%**	15.9%	12.6%	
교육부	1,727,923	26,095	3,459		1,491		43.1%
(비중)	7.7%	9.2%	9.2%	2.8%	5.7%	2.3%	
농촌진흥청	622,985	14,367	713		313		43.9%
(비중)	2.8%	1.9%	5.0%	0.6%	2.2%	0.5%	
해양수산부	599,569	13,656	1,000		347		34.7%
(비중)	2.7%	1.8%	7.3%	0.8%	2.5%	0.5%	
보건복지부	634,453	12,208	965		476		49.3%
(비중)	2.8%	1.6%	7.9%	0.8%	3.9%	0.7%	
총계	22,427,895	750,965	122,050		65,104		**53.3%**
(비중)	100%	100%	16.3%	100%	8.7%	100%	

- [Table 2] Comparison of the weight of research expenses by ministry

To find out the degree of SW convergence among different departments, the proportion of SW convergence R&D in the sample research funds of different departments was identified. In the case of the top five ministries with high research costs, 26.6% of SMB Departments are SW convergence tasks including SW convergence R&D activities. And The Ministry of Science Technology Information And Communication found that 18.4%, The Ministry of Trade Industry And Energy is charged 15.6 %, The Education Ministry was 9.2 %, and The Rural Development Administration 5.0 % were R&D tasks, including SW activities**.**

2) Current Status by Standard Classification of Science and Technology



(단위 : %, 백만 원)

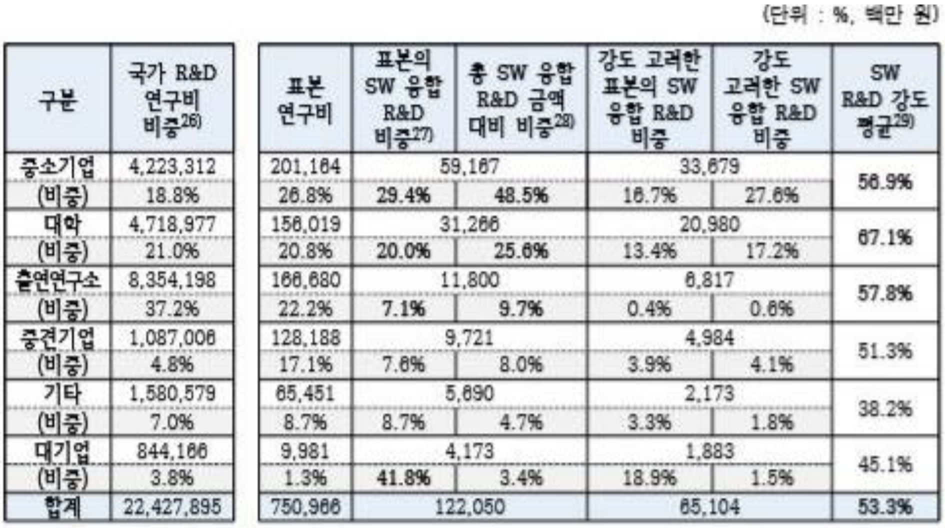
구분	국가 R&D 연구비 비중[22]	표본 과제 수	표본 연구비	표본의 SW 융합 R&D 비중[23]	SW 융합 R&D 비중[24]	강도 고려한 표본의 SW 융합 R&D 비중	강도 고려한 SW 융합 R&D 비중	SW R&D 강도 평균[25]
정보/통신	2,270,149	168	82,851	47,156		31,081		65.9%
(비중)	10.1%		11%	56.9%	38.6%	37.5%	47.7%	
보건의료	1,877,333	318	92,606	22,883		9,911		43.3%
(비중)	8.4%		12.3%	24.7%	18.7%	10.7%	15.2%	
기계	4,002,786	202	199,891	13,941		5,272		37.8%
(비중)	17.8%		26.6%	7.0%	11.4%	2.6%	8.1%	
건설/교통	1,021,558	72	26,672	11,086		6,506		58.7%
(비중)	4.6%		3.6%	41.6%	9.1%	24.4%	10.0%	
전기/전자	1,977,888	141	55,151	6,241		2,019		32.4%
(비중)	8.8%		7.3%	11.3%	5.1%	3.7%	3.1%	
...						...		
농림수산식품	1,324,292	326	40,086	1,573		570		36.2%
(비중)	5.9%		5%	3.9%	1.3%	1.4%	0.9%	
에너지/자원	1,143,429	65	23,066	418		256		61.2%
(비중)	5.1%		3.1%	1.8%	0.3%	1.1%	0.4%	
환경	560,545	63	19,998	1,330		603		45.3%
(비중)	2.5%		2.7%	6.6%	1.1%	3.0%	0.9%	
계	-	2005	750,966	-	122,050	-	65,104	53.3%

- [Table 3] Comparison of research costs by national science and technology standard classification (National R&D vs. SW convergence R&D)

In conclusion, Information/communication is not only subject to more SW-convergence R&D studies than other areas, but also to the task of SW-converged R&D. Additionally, it can be seen that there is a lot of SW convergence research being carried out in areas other than information/communication, in construction/transportation, health care and electricity/electronics.

On the other hand, in the Environmental, Agricultural, Forestry and Fisheries Food, Energy/Resource, Nuclear energy and Media /Communication /Civil Information field, fewer SW-converged R&D are actually performed in a variety of fields.

3) **A research agent**



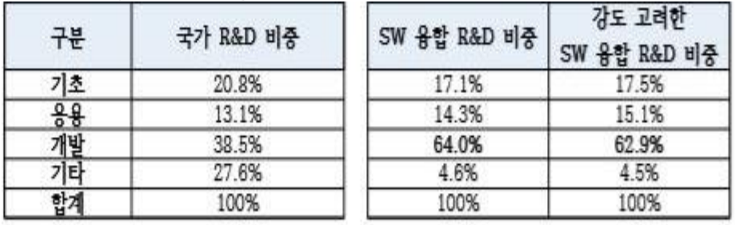
(단위 : %, 백만 원)

구분	국가 R&D 연구비 비중[26]	표본 연구비	표본의 SW 융합 R&D 비중[27]	총 SW 융합 R&D 금액 대비 비중[28]	강도 고려한 표본의 SW 융합 R&D 비중	강도 고려한 SW 융합 R&D 비중	SW R&D 강도 평균[29]
중소기업	4,223,312	201,164	59,167		33,679		56.9%
(비중)	18.8%	26.8%	29.4%	48.5%	16.7%	27.6%	
대학	4,718,977	156,019	31,266		20,980		67.1%
(비중)	21.0%	20.8%	20.0%	25.6%	13.4%	17.2%	
출연연구소	8,354,198	166,680	11,800		6,817		57.8%
(비중)	37.2%	22.2%	7.1%	9.7%	0.4%	0.6%	
중견기업	1,087,006	128,188	9,721		4,984		51.3%
(비중)	4.8%	17.1%	7.6%	8.0%	3.9%	4.1%	
기타	1,580,579	65,451	5,690		2,173		38.2%
(비중)	7.0%	8.7%	8.7%	4.7%	3.3%	1.8%	
대기업	844,166	9,981	4,173		1,883		45.1%
(비중)	3.8%	1.3%	41.8%	3.4%	18.9%	1.5%	
합계	22,427,895	750,966	122,050		65,104		53.3%

– [Table 4] Comparison of the proportion of research costs by subject
(National R&D vs. SW Convergence R&D)

If you look at the top three performing entities with high sample research costs, It was found that 29.4% of small and medium enterprises, 20% of universities, and 7.1% of research institutes performed tasks that included SW convergence R&D activities Of the three individuals with the highest sample research funds, Small and Medium enterprises (29.4%) are the ones with the most active SW convergence. The intensity of SW-converged R&D (56.9 percent) is also higher than the average (53.3 percent).

4) Status by R&D stage

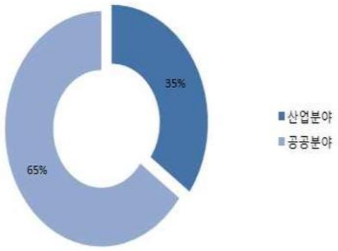


구분	국가 R&D 비중	SW 융합 R&D 비중	강도 고려한 SW 융합 R&D 비중
기초	20.8%	17.1%	17.5%
응용	13.1%	14.3%	15.1%
개발	38.5%	64.0%	62.9%
기타	27.6%	4.6%	4.5%
합계	100%	100%	100%

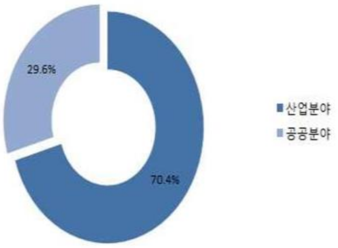
– [Table 5] Based on research costs, specific gravity
(National R&D vs. SW convergence R&D)

It was found that the percentage of total research cost considering the step-by-step strength of SW convergence R&D was 62.9% for development research, 17.5% for basic research and 15.1% for applied research.

5) Application field



[Figure 1] Percentage of national R&D

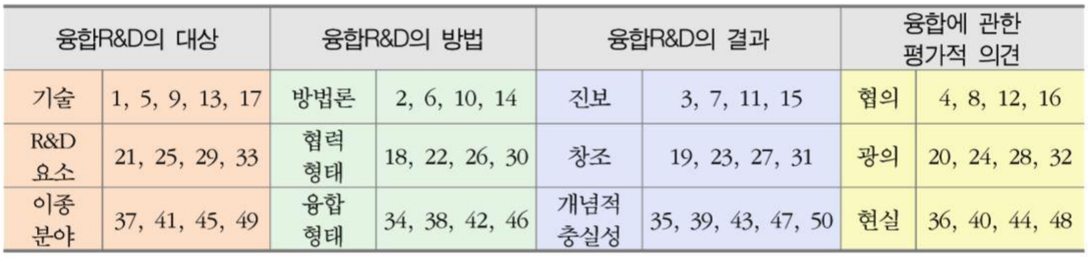


[Figure 2] Percentage of SW Convergence R&D

[Figure] and [Figure 2] show that SW convergence R&D is often used in industry rather than in the public sector. So we can see that there is a great need to actively reflect the demand of industry in the performance process of SW convergence R&D.

출처: 2019. 3. 6. 제2018-014호 융합 R&D 현황 분석 및 시사점- 서영희, 공영일

2.2 A Study on the Different Interests' Perception of Convergence R&D through Q Analysis



융합R&D의 대상		융합R&D의 방법		융합R&D의 결과		융합에 관한 평가적 의견	
기술	1, 5, 9, 13, 17	방법론	2, 6, 10, 14	진보	3, 7, 11, 15	협의	4, 8, 12, 16
R&D 요소	21, 25, 29, 33	협력 형태	18, 22, 26, 30	창조	19, 23, 27, 31	광의	20, 24, 28, 32
이종 분야	37, 41, 45, 49	융합 형태	34, 38, 42, 46	개념적 충실성	35, 39, 43, 47, 50	현실	36, 40, 44, 48

- [Table 6] Basic structure and unique number of Q samples

＊ The number in the table is the unique number of the Q sample used in the actual survey (true statement card)

A small number of people are recruited and classified according to [condition 1].

[Condition 1]

A person of special interest in the subject

A person who can give an impartial opinion

Authority or expert on the subject

A person of general interest

A person who is obscure or uninterested in the subject.

The purpose of this study is to analyze the results of the evaluation of the recruited persons in depth to find various opinions and to seek interpretative understanding accordingly. The above study conducted interviews with a total of 36 people with basic knowledge of Convergence R&D, Convergence R&D from Nov. 1 to Nov. 20, 2014.

As a result, the results of ◦ Convergence R&D were classified into four types of recognition that focused on results, 「Evaluation opinion, 」Fusion R&D method, 「Complex visualization of interdisciplinary fields, and 「Fusion R&D results and evaluative opinion. And each type of recognition has a common share.

Number 15 (result): Convergence R&D is to increase existing
Efficiency through convergence.
Number 14 (Method): Convergence in cooperation is centered on the leader.
And the more complex a fusion is, the more active a network of researchers is
Number 18 (Method): Convergence R&D consists of co-operating elements
That make up the convergence.
No. 26 (method): Convergence R&D consists of chemical bonds between
The elements that make up the convergence.
Number four (opposition): Convergence R&D refers to the chemical bond between
New technologies of different kinds.
Number 6 (Method): Convergence R&D is a way for various technologies/science
To interact and combine chemically.
No. 20 (opposition): The scope of convergence R&D continues to expand with
The development of technology and environmental changes.
No. 34 (method): Convergence R&D is a process of recombination
By existing components.
No. 16 (opposition): The phenomenon of increasing convergence R&D scope is
Not desirable for convergence R&D development..

The above opinions are common in the four types of opinions.

출처: 연구보고 2015-024, 융합 R&D 추진현황 분석 및 활성화 방안 - 김홍영, 박소희

3 Method

3.1 The Importance of Data Visualization

According to Blotter & Media, 'Capacity to utilize data' can be divided into 'data analysis' and 'visual storytelling'. The former refers to the ability required for the process from the stage of data processing to the stage of data processing, which technically collects and purifies data, to the analysis of data using analytical techniques. The latter is the ability to tell stories by visualizing the results of the data analysis. In the past, the boundary between these two capabilities was clear. However, the recent use of 'data visualization' is making it important to be able to encompass the two capabilities. Data visualization allows data to be explored and story-telling using data, even without the technical expertise to handle data, and without the design capabilities for visual storytelling.

Data visualization 1) presents a large amount of data using visual elements, making it recognizable at a glance. 2) Even without expertise in data analysis, anyone can easily find data insight. 3) More accurate data analysis results than summary statistics can be obtained. Data visualization is used as a data navigation method for accurate analysis, as well as simply for the purpose of communicating analysis results. 4) Effective data Insight sharing enables data-based decision making. 5) There are numerous fields and methods to utilize data visualization. For these reasons, data visualization is very important.

3.2 Data Visualization Tools: Spitfire

Data visualization includes various tools such as vRik, Qlikview, Tableau, Spitfire, and more. In particular, this study will use Spitfire to visualize data. Spitfire is a data analytics and visualization solution and is used by more than 4,000 companies worldwide. Data is extracted from DB, visualization charts are prepared, and core statistical analysis functions such as regression, clustering, and ANOVA are built in. Not only can traditional charts such as the Scatter, Line, and Pie Chart and so on be much richer, but they can also utilize advanced visualization charts such as Tree Map, Heat map, etc. By automatically filtering the entered data columns according to the data attributes, real-time inter-analysis and selective navigation between data sets can be made, and more fundamental in-depth analysis can be done by drill-down analysis, allowing analysis results to be freely "narrow but deeper" based on various data attributes..

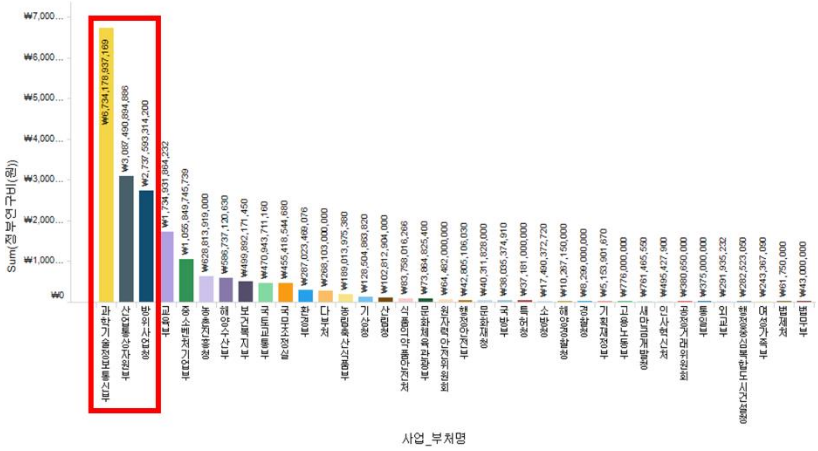
1 ANALYSIS

4.1 Basic analysiss

In the national R&D project carried out in 2017, a total of 19,392,668,129,240, government research funds of 61,280 data were used. Data on these were analyzed in four categories: ministries, research subjects and regions, applications and economic and social objectives and technologies.

1) Scale of government research funds by project department

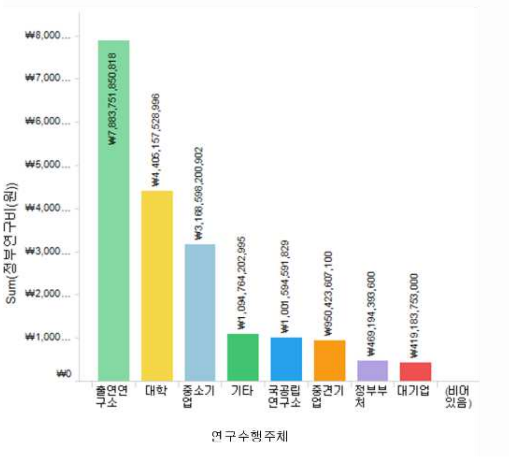
In order to find out the government research funds for each project department, the proportion of the sample research costs for each department was verified by visualizing them on a bar graph. For the top three ministries with high research funds for samples, the Ministry of Science, Technology and Information and Communication was found to be 34.7 percent at 6,734,178,937,169 won, 15.8 percent at the Ministry of Industrial and Transportation, and 14.1 percent at the Defense Acquisition Program Administration, 2,737,593,314,200 won. The top three ministries account for 64.7 percent of the total research funds.



[chart 1-1]Government research expenses according to business department name

2) Scale of government research funds by research subjects

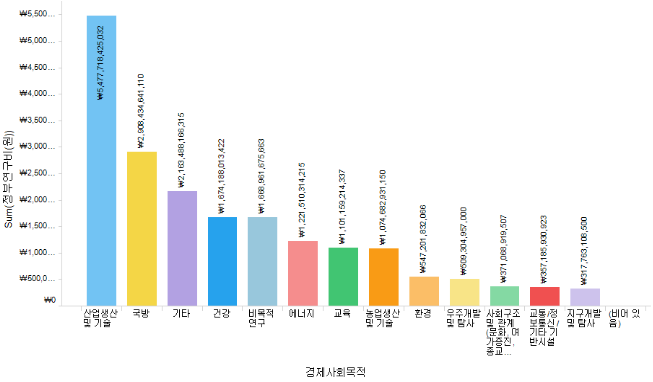
We looked into government research funds for each business department. In the case of the top three research subjects with high research costs, 40.7 percent of the respondents were found to be 7,883,751,850,818 won, 15.8 percent of the universities, 4,405,157,528,996 won, and 14.1 percent of small and medium enterprises, 3,168,598,200,902. The top three ministries account for 64.7 percent of the total research funds. In addition, the government's research funds in Daejeon, where there are many research institutes, were the largest, followed by the Seoul Metropolitan Government, which has many universities and industrial institutes.



[chart 1-2] Government research expenses according to research subjects

3) Scale of government research expenses by economic and social purpose

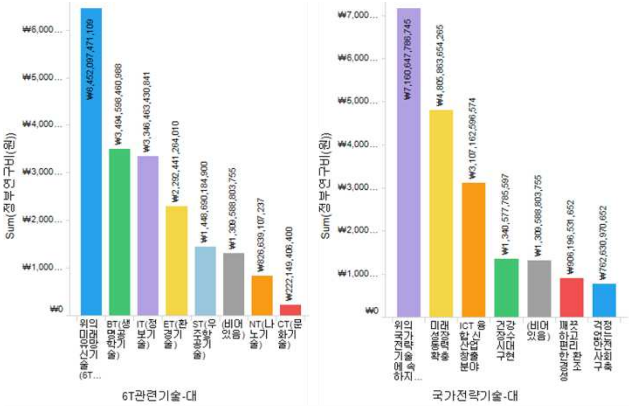
Government research funds were identified for each application area and purpose. For the top three application areas with high research costs, it was found that 18.0% of the total applied areas were CU3,494,588, 17.3% of IT (Information Technology), CU3,430,841 and 11.8% of ET (Environmental Technology) were KRW 2,292,441,264,0100. The top three ministries account for 48.1 percent of the total research funds. For the top three national strategic technologies with high research costs, 24.8 percent of the future growth engines will be increased to 4,805,863,654,265 won, 16.0 percent, 3,107,162,596 won,574 won, and 6.9 percent, respectively, in the ICT convergence new industry creation sector, and 1,340,577,785 and 597. The top three ministries account for 47.7 percent of the total research funds.



[chart 1-3]적용분야에 따른 정부연구비]

4) Scale of government research costs by technology

We looked at government research costs for each technology. Technology is divided into 6T-related technologies and national strategic technologies. For the top three 6T-related technologies with high research costs, it was found that BT (Bioengineering technology) was 18.0%, 3,494,598,463,460 and988 won, IT (information technology) was 17.3%, 3,346,463,430,841 won, and ET (environmental technology) was 2,292,441,264,010.10 won. The top three ministries account for 48.1 percent of the total research funds. For the top three national strategic technologies with high research costs, 24.8 percent of the future growth engines will be increased to 4,805,863,654,265 won, 16.0 percent, 3,107,162,596 won,574 won, and 6.9 percent, respectively, in the field of ICT convergence new industry creation, and 1,340,577,785597 won, respectively. The top three ministries account for 47.7 percent of the total research funds.



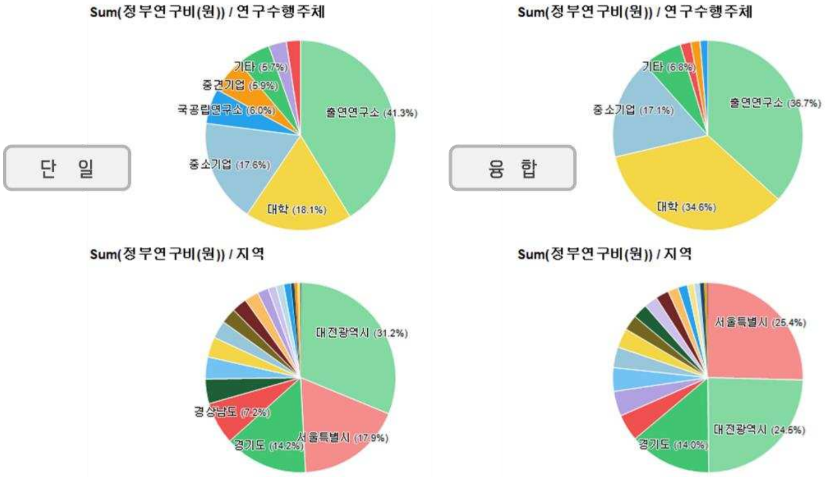
[chart 1-4] Government research expenses according to business department name

4.2 Comparison with Converged R&D

The national R&D projects were classified according to the task classification. As a result, we have recognized the characteristics of universities in the convergence R&D business. Looking at the proportion of the total data, the single task was 15,505,301,793,904 won respectively, and the convergence task was 13.3%, accounting for 1,309,588,803,755 and Category 0 was analyzed as a humanities project.

1) Scale of government research funds for single and convergence tasks by research subjects and regions

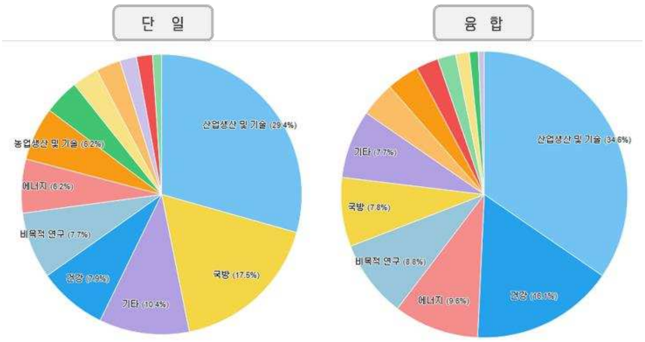
We looked at the differences between the ratio of the single task subject to the task classification and the convergence task to the research subject. Both single and convergence tasks were found to be the top performing institute, the second-largest university and the third-largest small and medium-sized enterprises. The biggest difference among them was "college," which accounted for 18.1 percent of the total, but nearly doubled to 34.6 percent for the convergence project. As a result, the Seoul Metropolitan Government, which has the largest number of universities, increased from 17.9 percent to 25.4 percent.



[chart 2-1] Government research costs for single and convergence tasks according to research subjects and regions

2) Size of government research funds for single and converged tasks by economic and social purpose

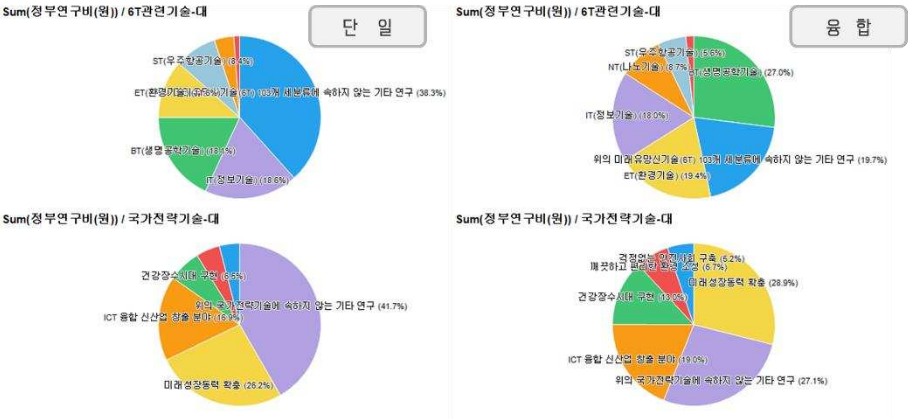
'Industrial production and technology' (29.4% and 34.6% respectively) accounted for the largest share of both single and converged tasks. Following industrial production and technology, national defense (17.5 percent) accounts for a large percentage of the single task, and health (16.1 percent) for the convergence task.



[chart 2-2] Government's Research on Single and Convergence Tasks by Economic and Social Objectives

3) Scale of government research costs for single and convergence tasks by technology

By technology, it was divided into 6T-related technologies and national strategic technologies. If we look at 6T-related technologies first, IT (Information Technology) accounted for the largest portion of the single task (except for studies not belonging to other_related technologies) with 18.6%. It can be found that BT (biotech technology) accounts for the largest portion of the convergence task at 27%, and ET (environmental technology) is also a larger proportion than a single task. In terms of national strategic technology, both single and convergence tasks are heavily invested in expanding future growth engines, and 13 percent of convergence tasks are implemented in the era of health longevity, accounting for twice the rate of a single task.



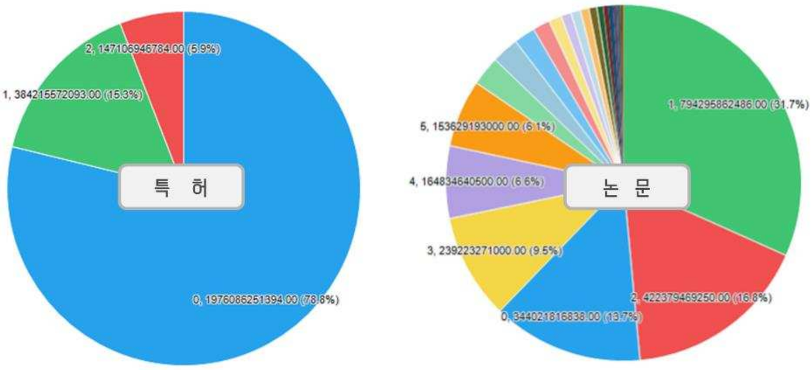
[chart 2-3] 6T Government research funds for single and convergence tasks according to the relevant technologies and national strategic technology

When analyzing a single task and a convergence task, compared to a convergence task, the single task is more expensive to spend on national defense, and thus the proportion is large enough for the Defense Acquisition Program Administration. On the other hand, convergence projects are considered to have a high percentage of investments in health and life technologies, so we could see that Korea invests a lot in industries that combine bio and bio-technology.

However, in the light of the ambiguity of the criteria for dividing a convergence task and a single task, incorrect judgement may exist to analyse a converged R&D project separately from a single R&D project.

4.3 Analysis based on performance (papers, paten)

Outliers of three variables [government research expense, number of patents, number of papers] were removed through the 3 sigma method before performance analysis was performed. The standard was applied to 20 papers, up to three patents, and government research costs of $9,600,000,000. Afterwards, the investment standards will be analyzed through a paper and a patent to enhance the efficiency of the current national R&D projects.



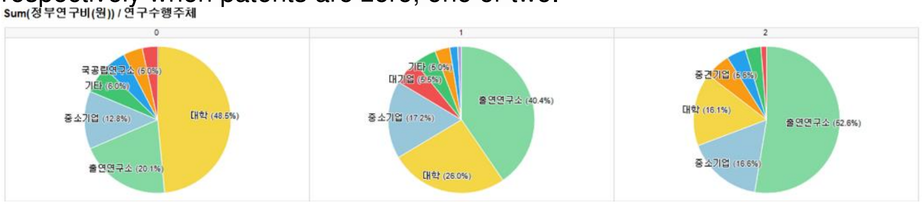
[chart 3-1] Percentage of government research funds based on the number of patents and papers

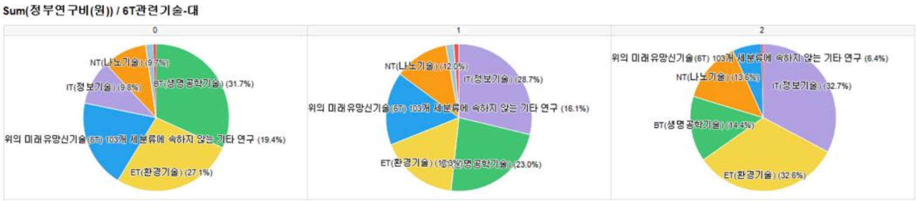
First of all, if you look at the size of government research funds based on the number of patents, 78.8 percent of them are without patents. The size of government research funds according to the number of papers accounted for 31 percent, 16.8 percent and 13.7 percent of cases where one is fruitless.

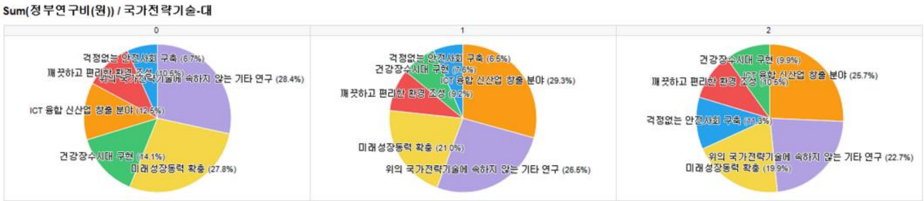
1) Percentage of government research expenses for each variable according to the number of patents

According to the government research funds by research subjects, 48.5 percent of universities and 20.1 percent of research institutes account for zero patents. However, if a single patent is granted, 40.4 percent of the research institute will participate in the program, and 26 percent of universities will take up the portion. In addition, when two patents are issued, 52.6 percent of the participating research institutes account for a larger portion, and 16.1 percent of the universities are about three times smaller than those without patents.

In terms of technology, IT (Information Technology) and ICT convergence new industry creation sectors are large at 9.8%, 28.7%, and 32.7% respectively when patents are zero, one or two.





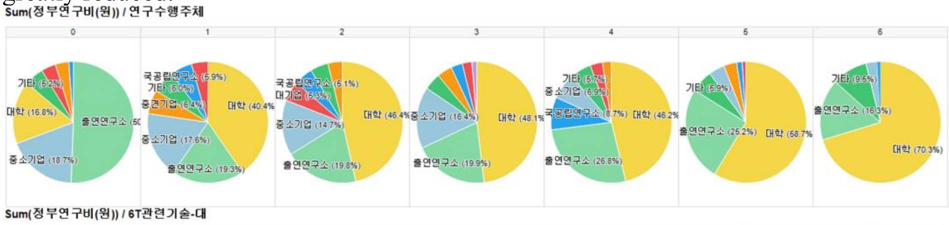


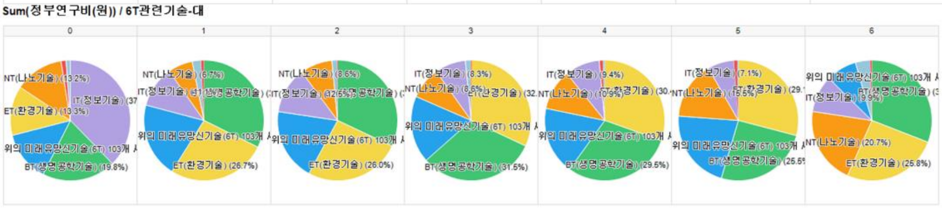
[chart 3-2] Percentage of government research funds for research subjects, 6T-related technologies, and national strategic technologies according to the number of patents

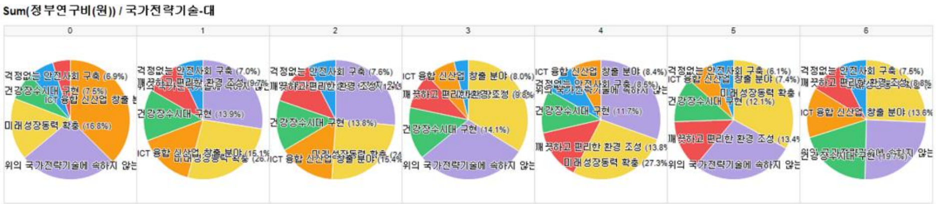
2) Percentage of government research costs for each variable according to the section of the thesis

According to the government research funds by research subjects, the more theses are published, the greater the portion of universities with 16.8 percent and 70.3 percent, respectively, when the sections of the papers are zero and six days old. On the other hand, 50.2 percent and 16.3 percent of the research institutes are showing a smaller portion.

In terms of technology, BT (biotechnical engineering) is 19.8% for dissertation section 0 and 30% for section 7. The higher the number of papers, the greater the weight. In addition, the technology for implementing the age of health longevity is 7.5 percent and 19.7 percent, respectively, in proportion to the number of papers. On the contrary, IT (Information Technology) and ICT convergence new industries will be greatly reduced.







[chart 3-3] Percentage of government research funds for research subjects, 6T-related technologies, and national strategic technologies according to the number of patents.

5 CONCLUSION

The research identified the distribution and performance of government research funds invested in R&D projects through 61,280 data from national R&D projects. The first thing to do was to visualize the entire data and then compare it with the convergence R&D business. By doing so, the single task accounted for a large portion of the government's research funds invested in the Defense Acquisition Program Administration and the national defense compared to the convergence project. On the other hand, convergence tasks accounted for a relatively large share of health, or biotechnology. We could see that the convergence industry in Korea is investing heavily in the bio sector. However, there is a limit that the criteria for dividing a convergence task and a single task are not clear, which could lead to inaccurate judgment.

Subsequent analyses were conducted by adding papers and patent variables to identify the performance of national R&D projects. As of 2017, a lot of investments are being made in participating laboratories, especially in IT (Information) technology and ICT convergence new industries.

In the case of papers, the higher the number of papers, the greater the importance of the education industry, such as universities, considering the characteristics that include human resources development projects. Moreover, the higher the number of papers, the greater the weight of biotechnology (BT) technology, the more achievements are achieved with many government investments.
As a result, the government's investment direction and efficiency can be identified for the convergence industry. You can intuitively see how much investment you have in which field you have.

So far, we have understood the hidden meaning of the data through visualization. This analysis is derived solely from data visualization, not from other statistical techniques, but from data visualization. This suggests to us that one method of visualization is powerful and a powerful tool for intuition and judgment.

Suggestion of Marketing direction through analysis of delivery data

- Focusing on the amount of phone calls made in Seoul–

Ju-han Oh[1], Hyeong-seok Kim[2], Seo-yeon Jeong[2],
Adviser: Prof. Wan-seop Joe

[1] Department of Manangemnt Information System, Chungbuk National University,
Cheongju, South Korea
[2] School of Business, Chungbuk National University,
Cheongju, South Korea
{ojh6280@naver.com}@naver.com, {khsean12@naver.com}@naver.com,
{jsy5669@naver.com}@naver.com

Contents

Table of Contents

Figure Table of Contents

Chapter 1. Introduction

1.1 Purpose and Background of Research

In the 2000s, the restaurant industry market doubled to about 128 trillion won in 2017 from about 64 trillion won in 2008 due to increased interest and demand in the restaurant industry due to the rise in income levels, the increase in the number of women's economic activities and the trend toward to the nuclear family. Also, as dining out becomes more common and mobile communication technology develops, the size of the food delivery market for all of 2018 reached 20 trillion won, according to industry estimates. The long-held delivery service, limited to a few sectors such as pizza, chicken, jjajangmyeon and napa wraps with pork(bossam), has recently expanded its scope, including snacks, raw fish, Japanese cuisine, the small intestines of cattle and desserts, with the advent of delivery applications in 2010, and the number of people using delivery applications has increased significantly from 870,000 in 2013 to 25 million in 2018. The volume of transactions through delivery applications also grew rapidly from 334.7 billion won in 2013 to 3 trillion won in 2018 after surpassing the 1 trillion won mark in 2015.

As a result, the competition is likely to get fiercer, with delivery applications that were the only one ‘national of delivery’ in 2010 - Yogiyo, the franchise's own app in addition to delivery bins - and more recently E-commerce companies such as Coupang and Wemap(Wemakeprice), as well as IT dinosaur companies Kakao and Naver announced their entry into the industry.

Therefore, we aim to analyze various variables and order call volume data, excluding promotion effects, to find marketing information that can be used by delivery application companies, franchisees, and individual operators. We will also finally set up a new virtual franchise and use the information we find to show the direction of our marketing strategy.

1.2 Data Source

Data including order information:

- Jun.~Aug., 2017, Seoul, Amount of calls made to the Chinese food industry (SKT Data Hub)
- Jun.~Aug., 2017, Seoul, Amount of calls made to the pizza industry (SKT Data Hub)
- Jun.~Aug., 2017, Seoul, Amount of calls made to the chicken industry (SKT Data Hub)
- Jun.~Aug., 2017, Seoul, Amount of calls made to the Chinese food industry (SKT Data Hub)
- Jun.~Aug., 2017, Seoul, Amount of calls made to the pizza industry (SKT Data Hub)
- Jun.~Aug., 2018, Seoul, Amount of calls made to the chicken industry (SKT Data Hub)

Data including order time, order place:

- Jun.~Aug., 2017, Seoul, Amount of calls made by delivery service sector (SKT Data Hub)
- Jun.~Aug., 2018, Seoul, Amount of calls made by delivery service sector (SKT Data Hub)

Seoul Latitude and Longitude Data

- Location Information for the Administrative Region of Seoul (Coordinate system: WGS1984) ,

(data.seoul.go.kr / Seoul open data square)

Meteorological data

- Longitudinal climate observation (climate data open portal)

1.3 Research method and Scope

This analysis utilizes the data contained in 1.2 after pretreatment and utilizes a program called Spotfire. The variables utilized in the analysis are the number of calls, industries, ordering time zones, administrative areas, age, gender, days of the week, highest temperature and average temperature, precipitation, rainfall, rainfall, tropical night occurrence or not, and sports events. Based on the number of calls, analyze one or two variables. The analysis methods used are a line graph, pie graph, bar graph, map chart, and correlation analysis. The scope of the study is limited to June-August data for 2017 and 2018 for efficient analysis.

Chapter 2. Theoretical Background and Data Analysis

2.1 Big Data

2.1.1 Definition of Big Data

Big data refers to a set of data with three-dimensional characteristics Volume, Velocity, Variety (3V). Big data can define all structured and unstructured data that is generated and circulated in real-time as big data, not just formal and meaningful words or numbers due to existing needs. Recently, there has been more discussion about big data and its definitions are becoming more diverse, either from a personal perspective or from a large number of corporate perspectives, as the purpose and meaning of the data being extracted through big data are different by analysts. As of 2011, IDC (International Data Corporation) announced that the amount of data generated in the last two years is much larger than that generated in the last 10 years and that the volume of digital information worldwide is doubling every two years. The difficulty of storage and processing due to the expansion of data size and type is also referred to as the era of big data, collectively referring to the trend in the modern information industry that storage and processing technologies can develop to analyze data and derive new information differently.

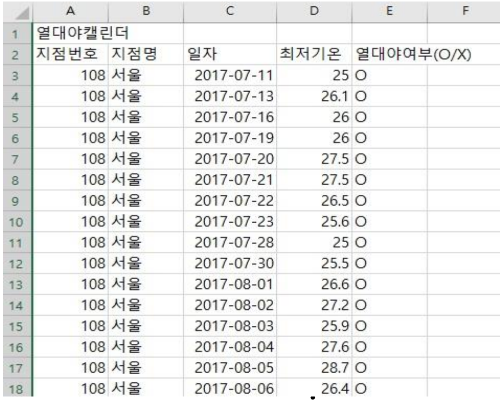
Traditional data analytics and big data analytics are also very different in their concepts, form and technology. Whereas conventional management information analysis analyzes data that has a high structured form, refined according to a certain form, big data analysis analyzes data that has a purified, unstructured, low structured form regardless of form. Analysis technology also uses social analysis methods, the latest statistical techniques, and artificial intelligence, including more advanced and complex distributed processing technologies, if previously simple distributed processing technology. New infrastructure technologies and analytical methods are being developed for these big data analytics, which are also called big data technologies in total.

2.1.2 Application of Big Data

Around the world, the government is paying attention to the value of big data and seeking ways to utilize big data in such areas as disasters, security, economy, health care, science and technology, transportation and government operations. In the U.S., big data is being used in various areas, including anti-terrorism, financial crime detection, big data collection, analysis and prediction systems for criminal DNA analysis, and disease distribution data by providing information on medicines. Britain and Australia have released information on their Internet homepages, providing a channel for people to use and participate in the information. Singapore is utilizing a system that collects and analyzes all national risks and seeks responses. The nation plans to establish a "national knowledge information platform" for utilizing big data to greatly expand the private opening of public knowledge information and boost the distribution of private knowledge information. The National Intelligence Strategy Committee's "Implementation of Smart Government Using Big Data" contains strategies for implementing information to lead the utilization of Big Data and infrastructure, and scenarios for using Big Data.

One of the areas that is blossoming the use of big data is the consumer industry. This is because we can effectively understand consumers by identifying social media content, etc., and by doing so, we can establish a more efficient marketing strategy. In addition, in the financial services sector, big data is often used to secure and maintain customers, detect financial fraud, and manage security risks. In the manufacturing and supply chain sector, big data analysis is also being used to optimize logistics, inventory and production lines, and big data is also playing an increasing role in identifying defects in manufacturing

2.2 The Preprocessing of Data.

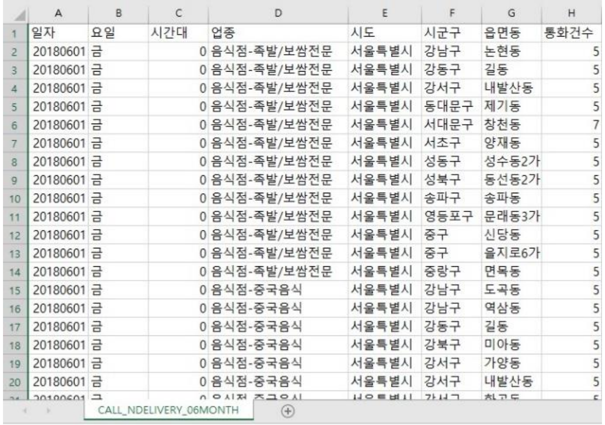


열대야캘린더

지점번호	지점명	일자	최저기온	열대야여부(O/X)
108	서울	2017-07-11	25	O
108	서울	2017-07-13	26.1	O
108	서울	2017-07-16	26	O
108	서울	2017-07-19	26	O
108	서울	2017-07-20	27.5	O
108	서울	2017-07-21	27.5	O
108	서울	2017-07-22	26.5	O
108	서울	2017-07-23	25.6	O
108	서울	2017-07-28	25	O
108	서울	2017-07-30	25.5	O
108	서울	2017-08-01	26.6	O
108	서울	2017-08-02	27.2	O
108	서울	2017-08-03	25.9	O
108	서울	2017-08-04	27.6	O
108	서울	2017-08-05	28.7	O
108	서울	2017-08-06	26.4	O

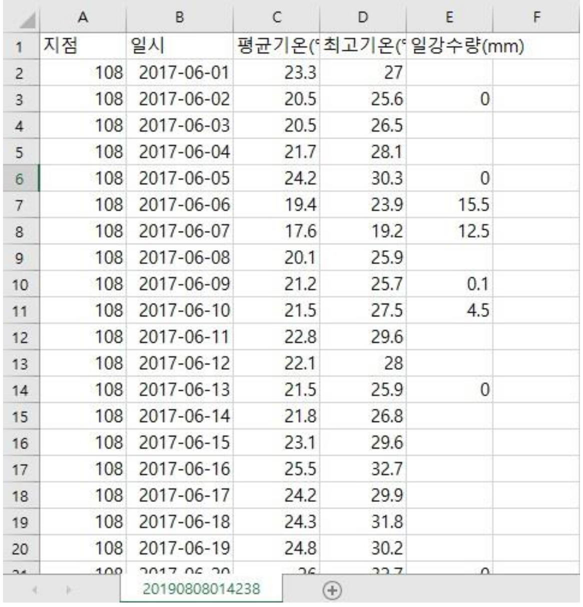
⋮

Table 1 [Tropical Night Data]



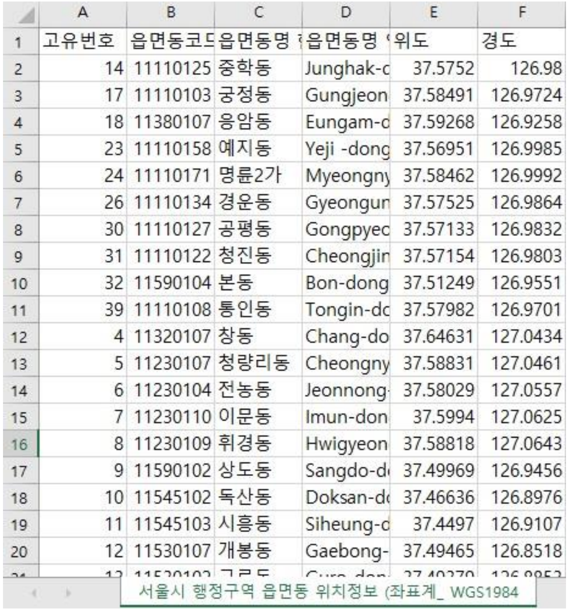
일자	요일	시간대	업종	시도	시군구	읍면동	통화건수
20180601	금	0	음식점-족발/보쌈전문	서울특별시	강남구	논현동	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	강동구	길동	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	강서구	내발산동	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	동대문구	제기동	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	서대문구	창천동	7
20180601	금	0	음식점-족발/보쌈전문	서울특별시	서초구	양재동	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	성동구	성수동2가	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	성북구	동선동2가	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	송파구	송파동	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	영등포구	문래동3가	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	중구	신당동	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	중구	을지로6가	5
20180601	금	0	음식점-족발/보쌈전문	서울특별시	중랑구	면목동	5
20180601	금	0	음식점-중국음식	서울특별시	강남구	도곡동	5
20180601	금	0	음식점-중국음식	서울특별시	강남구	역삼동	5
20180601	금	0	음식점-중국음식	서울특별시	강동구	길동	5
20180601	금	0	음식점-중국음식	서울특별시	강북구	미아동	5
20180601	금	0	음식점-중국음식	서울특별시	강서구	가양동	5
20180601	금	0	음식점-중국음식	서울특별시	강서구	내발산동	5
20180601	금	0	음식점-중국음식	서울특별시	강서구	화곡동	5

Table 2 [Data including order information]



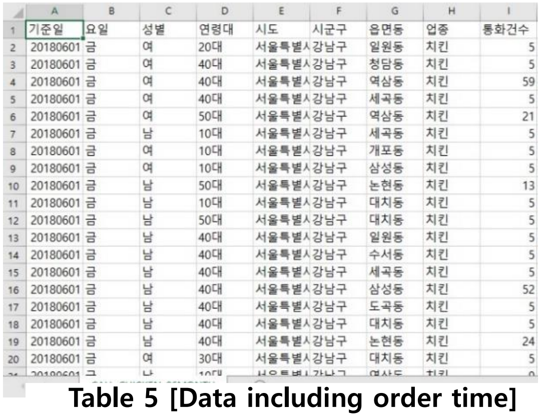
	A	B	C	D	E	F
1	지점	일시	평균기온(°	최고기온(°	일강수량(mm)	
2	108	2017-06-01	23.3	27		
3	108	2017-06-02	20.5	25.6	0	
4	108	2017-06-03	20.5	26.5		
5	108	2017-06-04	21.7	28.1		
6	108	2017-06-05	24.2	30.3	0	
7	108	2017-06-06	19.4	23.9	15.5	
8	108	2017-06-07	17.6	19.2	12.5	
9	108	2017-06-08	20.1	25.9		
10	108	2017-06-09	21.2	25.7	0.1	
11	108	2017-06-10	21.5	27.5	4.5	
12	108	2017-06-11	22.8	29.6		
13	108	2017-06-12	22.1	28		
14	108	2017-06-13	21.5	25.9	0	
15	108	2017-06-14	21.8	26.8		
16	108	2017-06-15	23.1	29.6		
17	108	2017-06-16	25.5	32.7		
18	108	2017-06-17	24.2	29.9		
19	108	2017-06-18	24.3	31.8		
20	108	2017-06-19	24.8	30.2		

Table 3 [Meteorological Data]



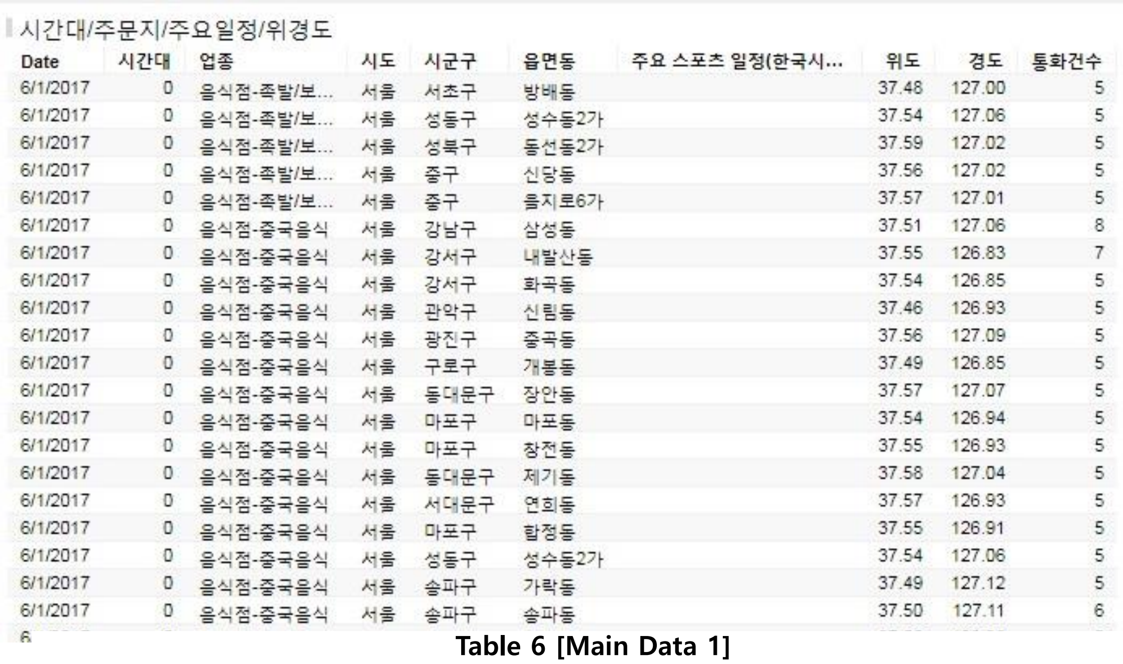
	A	B	C	D	E	F
1	고유번호	읍면동코드	읍면동명 (	읍면동명 (	위도	경도
2	14	11110125	중학동	Junghak-c	37.5752	126.98
3	17	11110103	궁정동	Gungjeon	37.58491	126.9724
4	18	11380107	응암동	Eungam-d	37.59268	126.9258
5	23	11110158	예지동	Yeji -dong	37.56951	126.9985
6	24	11110171	명륜2가	Myeongny	37.58462	126.9992
7	26	11110134	경운동	Gyeongun	37.57525	126.9864
8	30	11110127	공평동	Gongpyec	37.57133	126.9832
9	31	11110122	청진동	Cheongjin	37.57154	126.9803
10	32	11590104	본동	Bon-dong	37.51249	126.9551
11	39	11110108	통인동	Tongin-dc	37.57982	126.9701
12	4	11320107	창동	Chang-do	37.64631	127.0434
13	5	11230107	청량리동	Cheongny	37.58831	127.0461
14	6	11230104	전농동	Jeonnong	37.58029	127.0557
15	7	11230110	이문동	Imun-don	37.5994	127.0625
16	8	11230109	휘경동	Hwigyeon	37.58818	127.0643
17	9	11590102	상도동	Sangdo-d	37.49969	126.9456
18	10	11545102	독산동	Doksan-d	37.46636	126.8976
19	11	11545103	시흥동	Siheung-d	37.4497	126.9107
20	12	11530107	개봉동	Gaebong-	37.49465	126.8518

Table 4 [Seoul Administrative Region Location Data]



	A	B	C	D	E	F	G	H	I
1	기준일	요일	성별	연령대	시도	시군구	읍면동	업종	통화건수
2	20180601	금	여	20대	서울특별시	강남구	일원동	치킨	5
3	20180601	금	여	40대	서울특별시	강남구	청담동	치킨	5
4	20180601	금	여	40대	서울특별시	강남구	역삼동	치킨	59
5	20180601	금	여	40대	서울특별시	강남구	세곡동	치킨	5
6	20180601	금	여	50대	서울특별시	강남구	역삼동	치킨	21
7	20180601	금	남	10대	서울특별시	강남구	세곡동	치킨	5
8	20180601	금	여	10대	서울특별시	강남구	개포동	치킨	5
9	20180601	금	여	10대	서울특별시	강남구	삼성동	치킨	5
10	20180601	금	남	50대	서울특별시	강남구	논현동	치킨	13
11	20180601	금	남	10대	서울특별시	강남구	대치동	치킨	5
12	20180601	금	남	50대	서울특별시	강남구	대치동	치킨	5
13	20180601	금	남	40대	서울특별시	강남구	일원동	치킨	5
14	20180601	금	남	40대	서울특별시	강남구	수서동	치킨	5
15	20180601	금	남	40대	서울특별시	강남구	세곡동	치킨	5
16	20180601	금	남	40대	서울특별시	강남구	삼성동	치킨	52
17	20180601	금	남	40대	서울특별시	강남구	도곡동	치킨	5
18	20180601	금	남	40대	서울특별시	강남구	대치동	치킨	5
19	20180601	금	남	40대	서울특별시	강남구	논현동	치킨	24
20	20180601	금	여	30대	서울특별시	강남구	대치동	치킨	5

Table 5 [Data including order time]

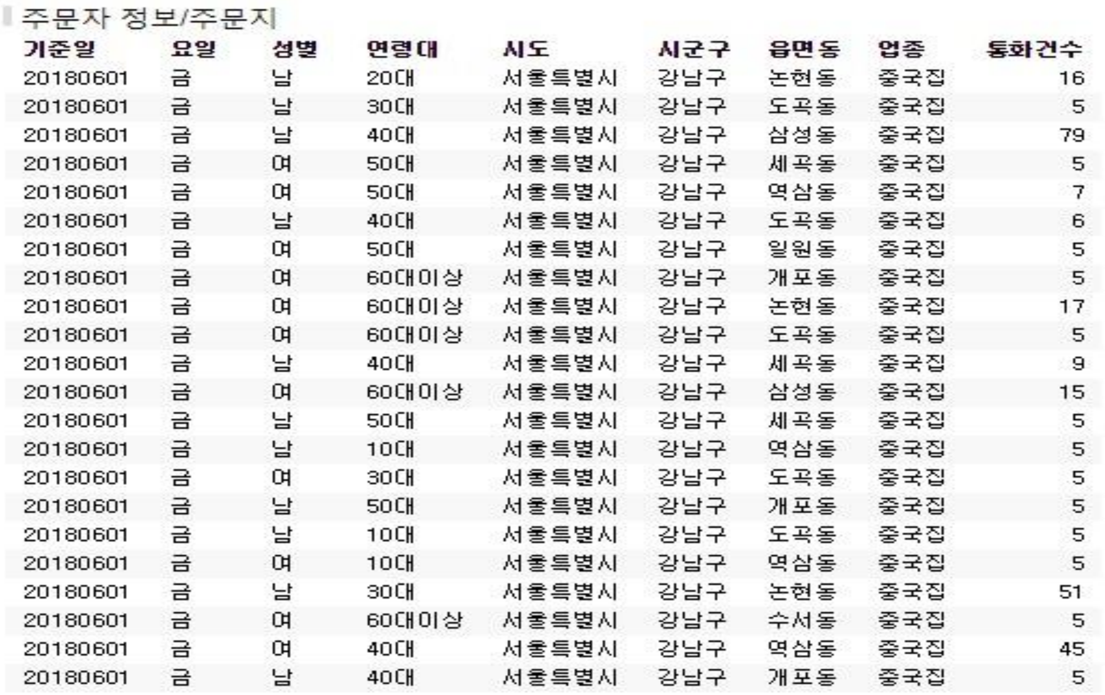


시간대/주문지/주요일정/위경도

Date	시간대	업종	시도	시군구	읍면동	주요 스포츠 일정(한국시...	위도	경도	통화건수
6/1/2017	0	음식점-족발/보...	서울	서초구	방배동		37.48	127.00	5
6/1/2017	0	음식점-족발/보...	서울	성동구	성수동2가		37.54	127.06	5
6/1/2017	0	음식점-족발/보...	서울	성북구	동선동2가		37.59	127.02	5
6/1/2017	0	음식점-족발/보...	서울	중구	신당동		37.56	127.02	5
6/1/2017	0	음식점-족발/보...	서울	중구	을지로6가		37.57	127.01	5
6/1/2017	0	음식점-중국음식	서울	강남구	삼성동		37.51	127.06	8
6/1/2017	0	음식점-중국음식	서울	강서구	내발산동		37.55	126.83	7
6/1/2017	0	음식점-중국음식	서울	강서구	화곡동		37.54	126.85	5
6/1/2017	0	음식점-중국음식	서울	관악구	신림동		37.46	126.93	5
6/1/2017	0	음식점-중국음식	서울	광진구	중곡동		37.56	127.09	5
6/1/2017	0	음식점-중국음식	서울	구로구	개봉동		37.49	126.85	5
6/1/2017	0	음식점-중국음식	서울	동대문구	장안동		37.57	127.07	5
6/1/2017	0	음식점-중국음식	서울	마포구	마포동		37.54	126.94	5
6/1/2017	0	음식점-중국음식	서울	마포구	창전동		37.55	126.93	5
6/1/2017	0	음식점-중국음식	서울	동대문구	제기동		37.58	127.04	5
6/1/2017	0	음식점-중국음식	서울	서대문구	연희동		37.57	126.93	5
6/1/2017	0	음식점-중국음식	서울	마포구	합정동		37.55	126.91	5
6/1/2017	0	음식점-중국음식	서울	성동구	성수동2가		37.54	127.06	5
6/1/2017	0	음식점-중국음식	서울	송파구	가락동		37.49	127.12	5
6/1/2017	0	음식점-중국음식	서울	송파구	송파동		37.50	127.11	6

Table 6 [Main Data 1]

In order to find out the impact of major sports schedules on the change of delivery call volume, we entered the specific date of the Asian Games, World Cup, etc. games and designated the same time zone information data and date as the same variable and joined in the left single-match, and the town name column of the location information data by Seoul administrative region and the town hall data including time zone were designated as the same variable and joined as the left single-match.

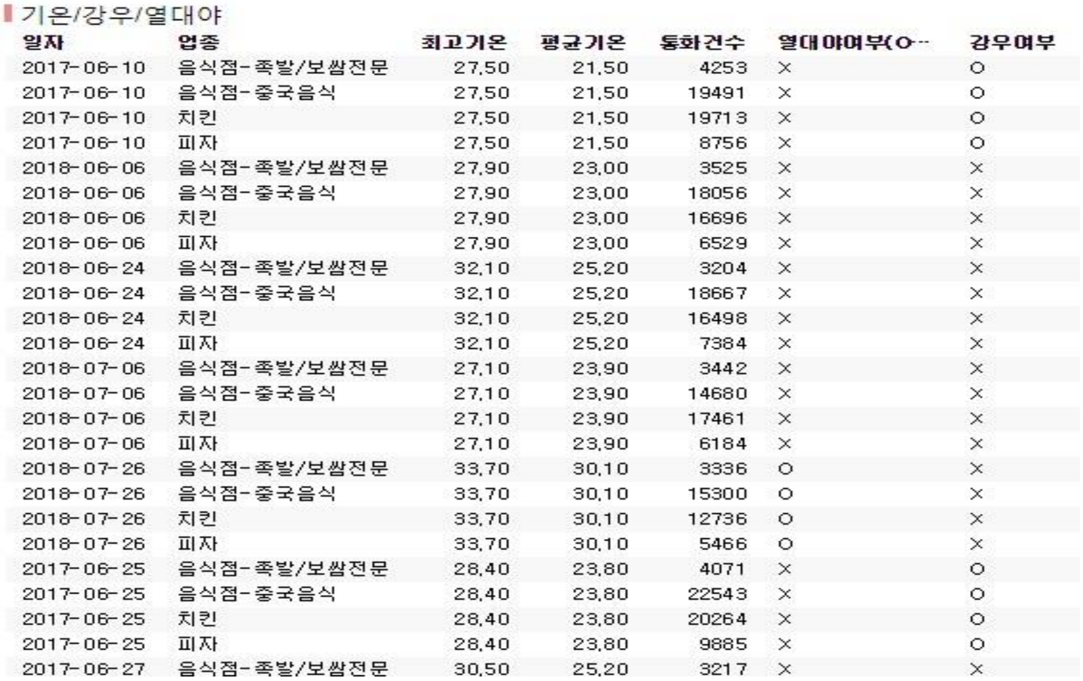


주문자 정보/주문지

기준일	요일	성별	연령대	시도	시군구	읍면동	업종	통화건수
20180601	금	남	20대	서울특별시	강남구	논현동	중국집	16
20180601	금	남	30대	서울특별시	강남구	도곡동	중국집	5
20180601	금	남	40대	서울특별시	강남구	삼성동	중국집	79
20180601	금	여	50대	서울특별시	강남구	세곡동	중국집	5
20180601	금	여	50대	서울특별시	강남구	역삼동	중국집	7
20180601	금	남	40대	서울특별시	강남구	도곡동	중국집	6
20180601	금	여	50대	서울특별시	강남구	일원동	중국집	5
20180601	금	여	60대이상	서울특별시	강남구	개포동	중국집	5
20180601	금	여	60대이상	서울특별시	강남구	논현동	중국집	17
20180601	금	여	60대이상	서울특별시	강남구	도곡동	중국집	5
20180601	금	남	40대	서울특별시	강남구	세곡동	중국집	9
20180601	금	여	60대이상	서울특별시	강남구	삼성동	중국집	15
20180601	금	남	50대	서울특별시	강남구	세곡동	중국집	5
20180601	금	남	10대	서울특별시	강남구	역삼동	중국집	5
20180601	금	여	30대	서울특별시	강남구	도곡동	중국집	5
20180601	금	남	50대	서울특별시	강남구	개포동	중국집	5
20180601	금	남	10대	서울특별시	강남구	도곡동	중국집	5
20180601	금	여	10대	서울특별시	강남구	역삼동	중국집	5
20180601	금	남	30대	서울특별시	강남구	논현동	중국집	51
20180601	금	여	60대이상	서울특별시	강남구	수서동	중국집	5
20180601	금	여	40대	서울특별시	강남구	역삼동	중국집	45
20180601	금	남	40대	서울특별시	강남구	개포동	중국집	5

Table 7 [Main Data 2]

Main data 2 did not go through the preprocessing.

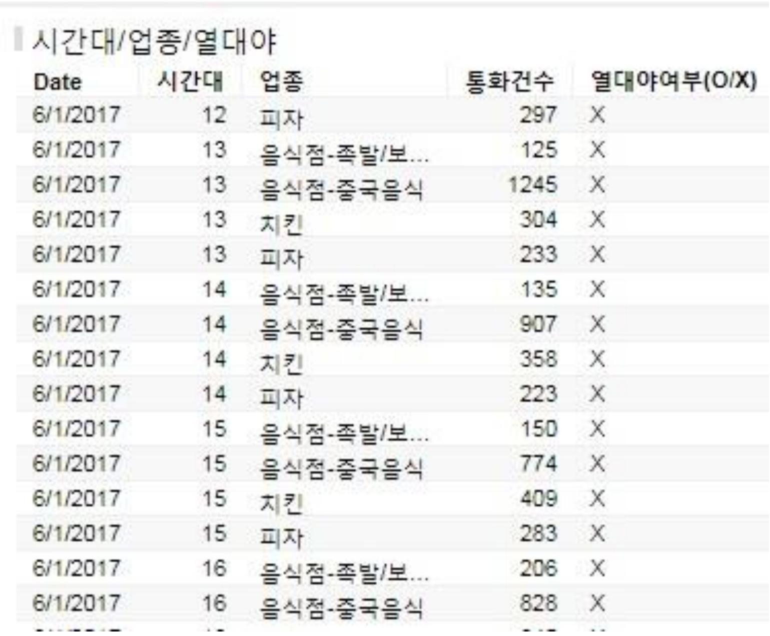


기온/강우/열대야

일자	업종	최고기온	평균기온	통화건수	열대야여부(O…	강우여부
2017-06-10	음식점-족발/보쌈전문	27.50	21.50	4253	×	○
2017-06-10	음식점-중국음식	27.50	21.50	19491	×	○
2017-06-10	치킨	27.50	21.50	19713	×	○
2017-06-10	피자	27.50	21.50	8756	×	○
2018-06-06	음식점-족발/보쌈전문	27.90	23.00	3525	×	×
2018-06-06	음식점-중국음식	27.90	23.00	18056	×	×
2018-06-06	치킨	27.90	23.00	16696	×	×
2018-06-06	피자	27.90	23.00	6529	×	×
2018-06-24	음식점-족발/보쌈전문	32.10	25.20	3204	×	×
2018-06-24	음식점-중국음식	32.10	25.20	18667	×	×
2018-06-24	치킨	32.10	25.20	16498	×	×
2018-06-24	피자	32.10	25.20	7384	×	×
2018-07-06	음식점-족발/보쌈전문	27.10	23.90	3442	×	×
2018-07-06	음식점-중국음식	27.10	23.90	14680	×	×
2018-07-06	치킨	27.10	23.90	17461	×	×
2018-07-06	피자	27.10	23.90	6184	×	×
2018-07-26	음식점-족발/보쌈전문	33.70	30.10	3336	○	×
2018-07-26	음식점-중국음식	33.70	30.10	15300	○	×
2018-07-26	치킨	33.70	30.10	12736	○	×
2018-07-26	피자	33.70	30.10	5466	○	×
2017-06-25	음식점-족발/보쌈전문	28.40	23.80	4071	×	○
2017-06-25	음식점-중국음식	28.40	23.80	22543	×	○
2017-06-25	치킨	28.40	23.80	20264	×	○
2017-06-25	피자	28.40	23.80	9885	×	○
2017-06-27	음식점-족발/보쌈전문	30.50	25.20	3217	×	×

Table 8 [Sub Data 1]

To find out the effect of weather conditions on the volume of ordered currency by industry, the highest/average temperature, tropical night conditions and rainfall date were included as a left single-match with the same variable. At this time, the time zone column was deleted through the pivot. This is because if the Time Zone column is not deleted, the average time zone of the weather condition is averaged instead of the daily mean of the weather condition. In addition, ordering information, which is a column that will not be used, was deleted.



시간대/업종/열대야

Date	시간대	업종	통화건수	열대야여부(O/X)
6/1/2017	12	피자	297	X
6/1/2017	13	음식점-족발/보...	125	X
6/1/2017	13	음식점-중국음식	1245	X
6/1/2017	13	치킨	304	X
6/1/2017	13	피자	233	X
6/1/2017	14	음식점-족발/보...	135	X
6/1/2017	14	음식점-중국음식	907	X
6/1/2017	14	치킨	358	X
6/1/2017	14	피자	223	X
6/1/2017	15	음식점-족발/보...	150	X
6/1/2017	15	음식점-중국음식	774	X
6/1/2017	15	치킨	409	X
6/1/2017	15	피자	283	X
6/1/2017	16	음식점-족발/보...	206	X
6/1/2017	16	음식점-중국음식	828	X

Table 9 [Sub Data 2]

Time zone inclusion data and tropical night data were left single-matched on a date basis and the Unnecessary Order Information column was deleted to see if there was any change in the volume of calls ordered during the night when tropical nights occurred.

2.3 Analysis

2.3.1 Call Volume by industry

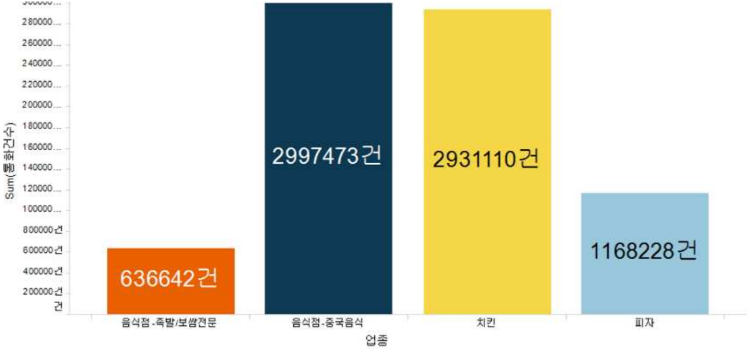


Figure 1 [Data including order time]

Based on data from June to August 2017 and 2018, Chinese food accounted for the largest number of cases among the four industries, with 2,997,473, followed by jokbal / napa wraps with pork(bossam) with 636,642 cases.

2.3.2 Order Calls Trend by Industry/Time

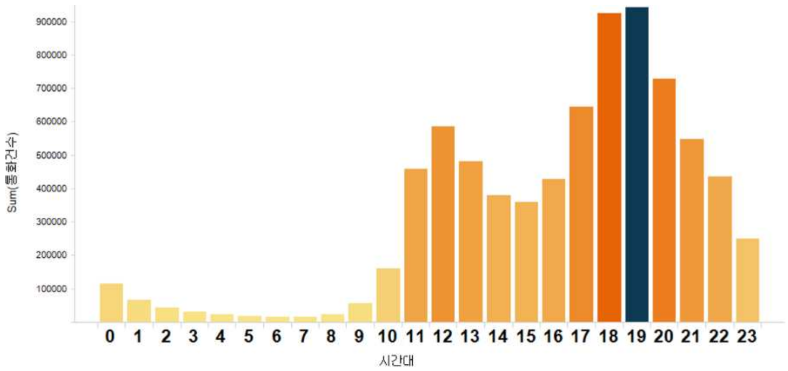


Figure 2 [Call volume by time zone of all four industry]

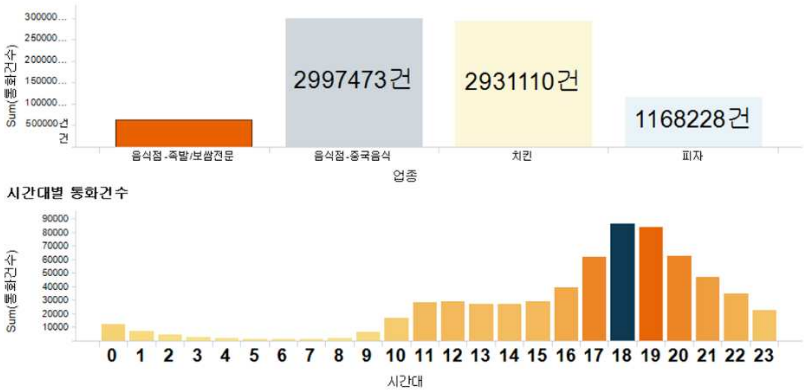


Figure 3 [Call volume by time zone of the jokbal/ bossam industry]

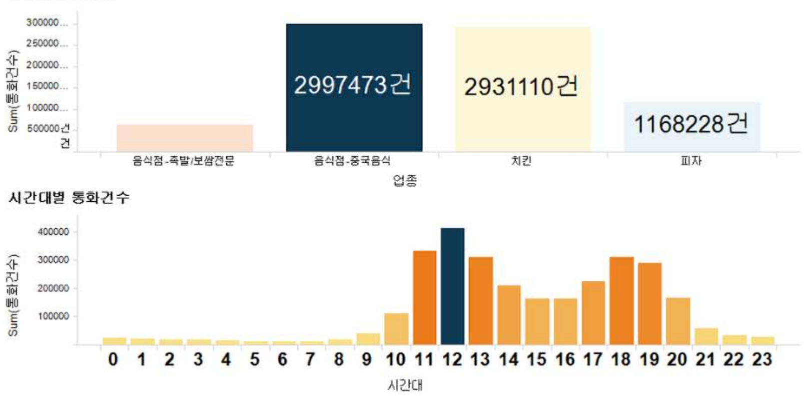


Figure 4 [Call volume by time zone of the chinese food industry]

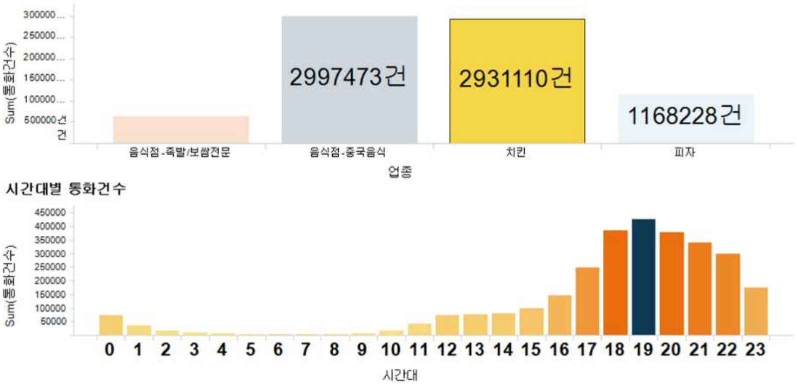


Figure 5 [Call volume by time zone of the chicken industry]

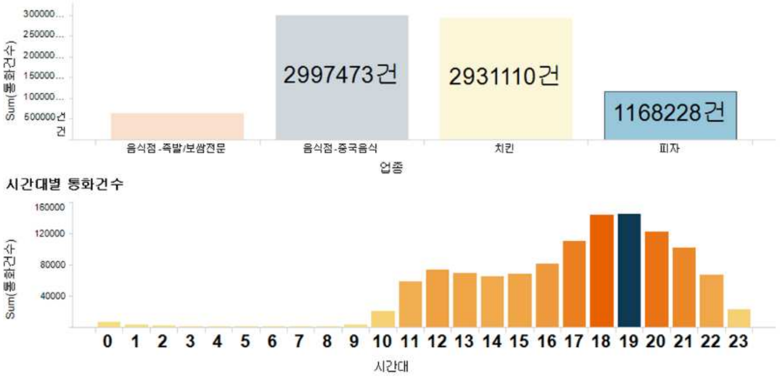


Figure 6 [Call volume by time zone of the pizza industry]

If you mark the entire industry on time, you'll find that you order the most food delivered at around 6 or 7 p.m. in the evening.

By industry, Chinese food has the highest volume of calls at about 400,000 at lunchtime, compared with orders from other industries being concentrated in the evening. In addition, in the case of chicken, the volume of calls from other industries, except Chinese restaurants, has reached its highest point in the evening and sharply decreased, while the figure is showing a relatively slow decline until 11 p.m.

2.3.3 Order Call Volume Trend by day of the week

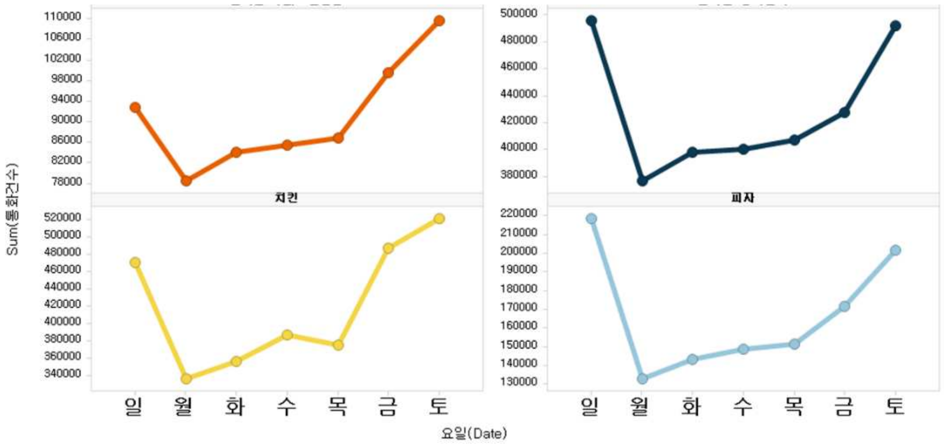


Figure 7 [Changes in call volume by industry/day]

If you look at the trends in the volume of ordered calls by day, it shows the lowest volume Monday and most calls are concentrated on the weekend. By taking advantage of this, related companies should positively consider promoting their products on weekends.

2.3.4 Comparison of Call Volumes by year

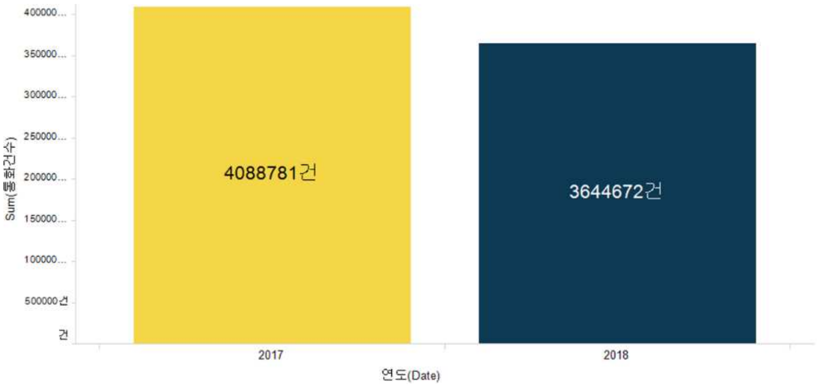


Figure 8 [Comparison of call volumes for 2017 and 2018]

The volume of calls decreased by about 10%, or 400,000 units, compared to orders made between June and August 2017. This seems to be due to an increase in public delivery APP usage. In addition, the current trend of declining currency volume is expected to continue in the future as many prominent companies can positively view the future delivery application market's outlook, with many promising companies signaling a market advance. Therefore, when launching new franchises or personal operators enter the restaurant business, they should actively consider partnerships with delivery application providers.

2.3.5 Analysis of Age and Gender

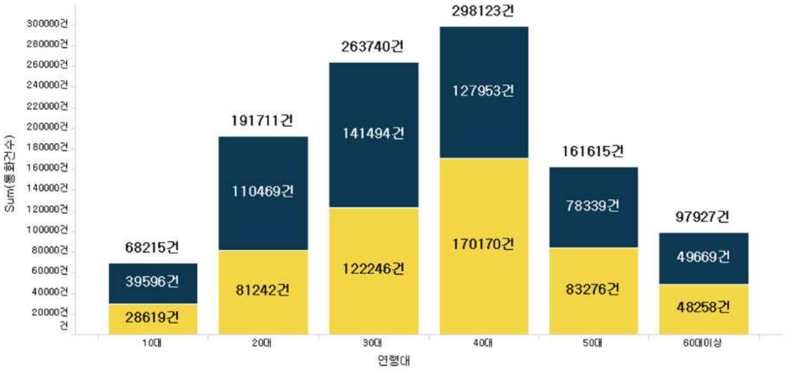


Figure 9 [Comparison of call volumes by age]

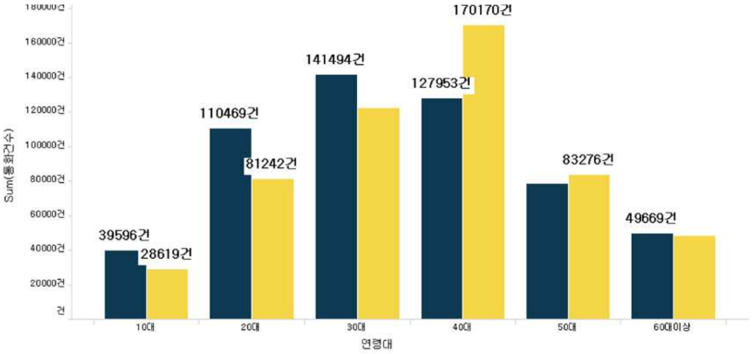


Figure 10 [Comparison of age/sex calls]

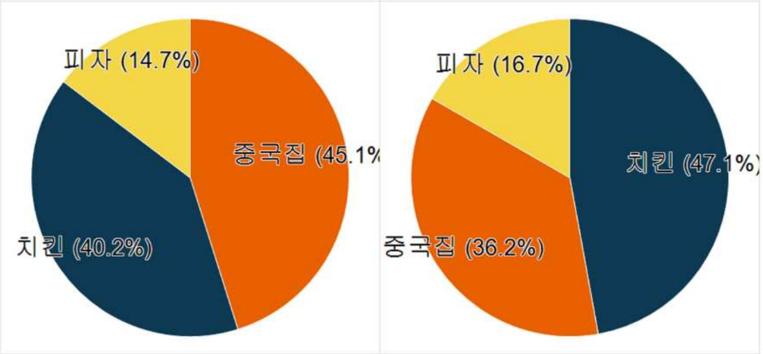


Figure 11 [Comparison of gender preferred industries (left: male, right: female)]

The group with the largest consumption power by age has a higher number of calls in their 30s and 40s compared to other age groups. Women in their 40s and men in their 30s were found to have the highest number of calls when gender and age groups were considered However, since the data used for analysis does not include orders through delivery applications, it cannot be concluded that they have the greatest consumption power. Men prefer Chinese food and women prefer chicken, the report showed.

2.3.6 Trends of Preferred Industries by Age

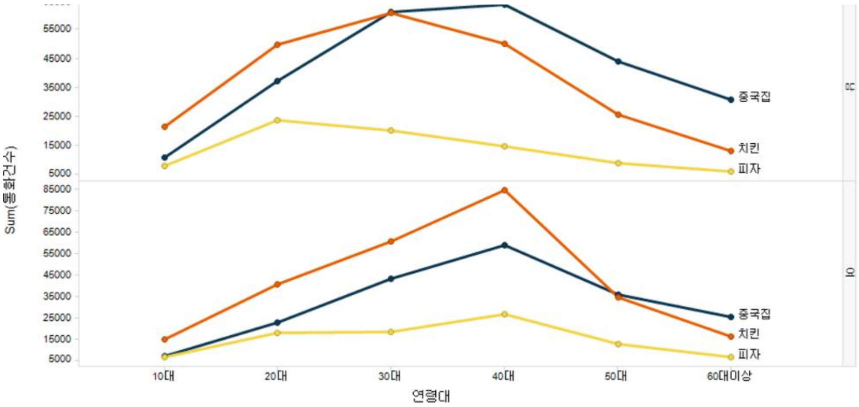


Figure 12 [Comparison of call volumes by age group/industry]

Line graphically to identify age-specific preferred industries. In men's case, Chinese restaurants outperform the amount of chicken calls based on their 30s, while women are in their 50s..

2.3.7 Comparison of Call Volumes by administrative region

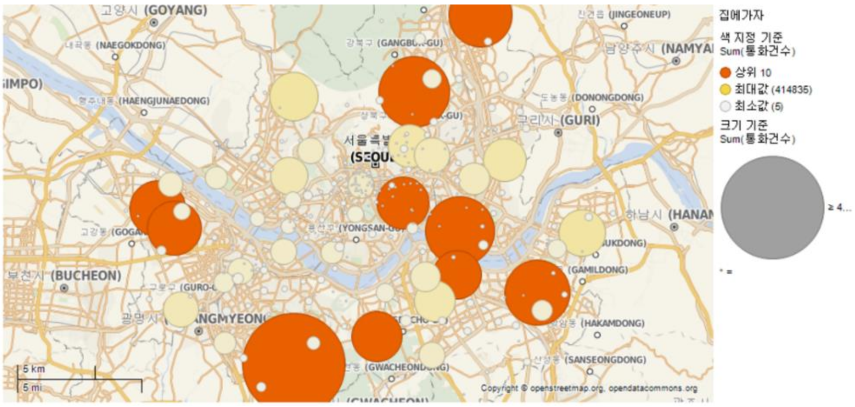


Figure 13 [Seoul City's Top 10]

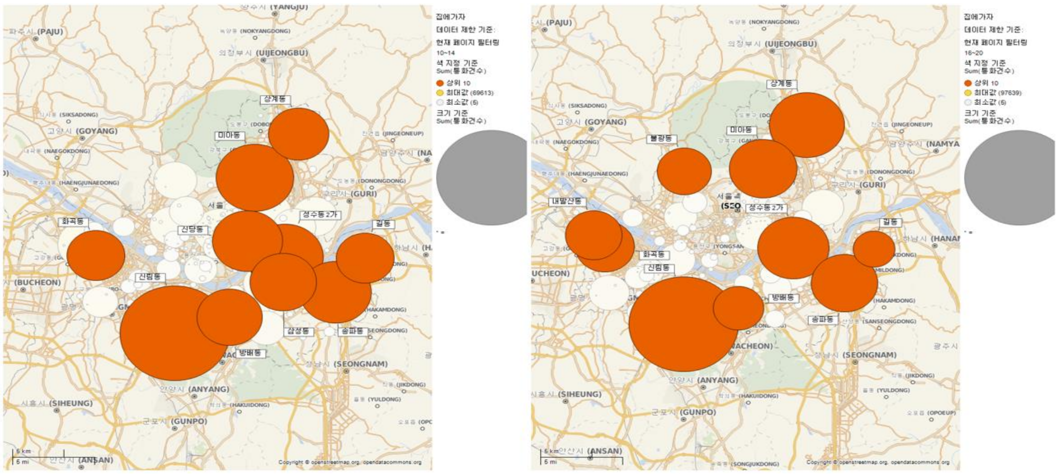


Figure 14 [Left: Day, Right: Night]

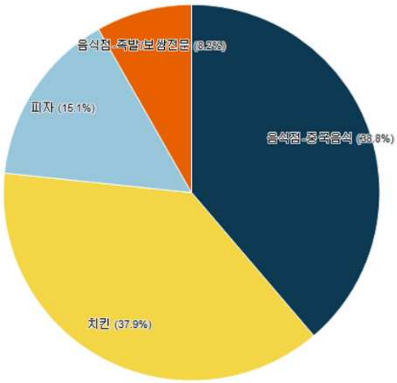


Figure 15 [Current Rate of Calls by Industry]

The areas expressed in orange on the map are the top 10 areas of call volume by administrative region. The map on the bottom left represents the top 10 areas of call volume in the daytime and the map on the right. Using pie charts, one can figure out the size of preferred industries by administrative region and their markets according to time zones, which can help select locations. The top 10 regions in the overall time zone are Sanggye-dong, Mia-dong, Sindang-dong, Seongsu-dong 2-ga, Songpa-dong, Samseong-dong, Bangbae-dong, Sillim-dong, Navalsan-dong and Hwagok-dong. The top 10 areas of the day zone (10-16 p.m.) are Sanggye-dong, Mia-dong, Seongsu-dong 2-ga, Sindang-dong, Gil-dong, Songpa-dong, Samseong-dong, Bangbae-dong, Shillim-dong, Hwagok-dong, Sanggye-dong, Bulkwang-dong, Seongsu-dong 2-ga, Songpa-dong, and Shindong-dong.

2.3.8 Analysis of the Change of the Amount of Money According to the Weather Condition

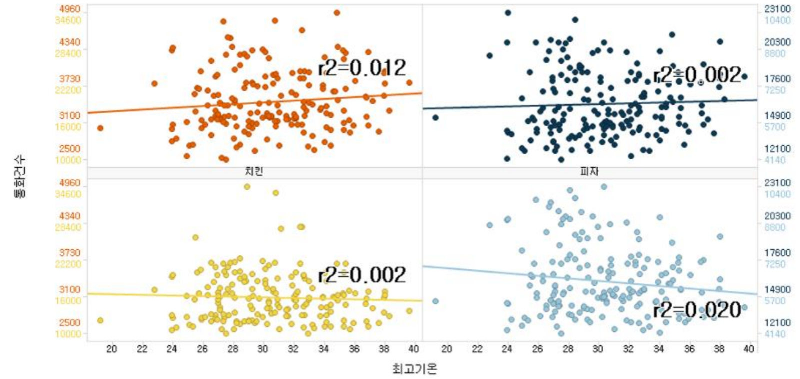


Figure 16 [Analyze the correlation of peak temperature/currency volume]

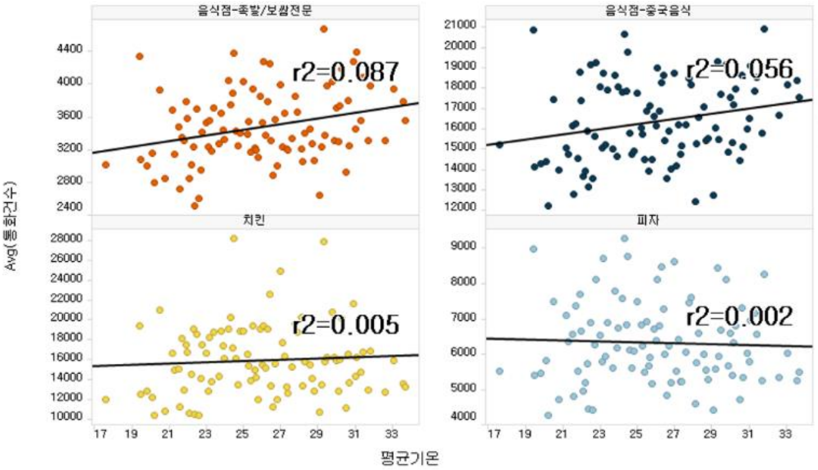


Figure 17 [Analysis of Correlation between Average Temperatures and Currency Volume]

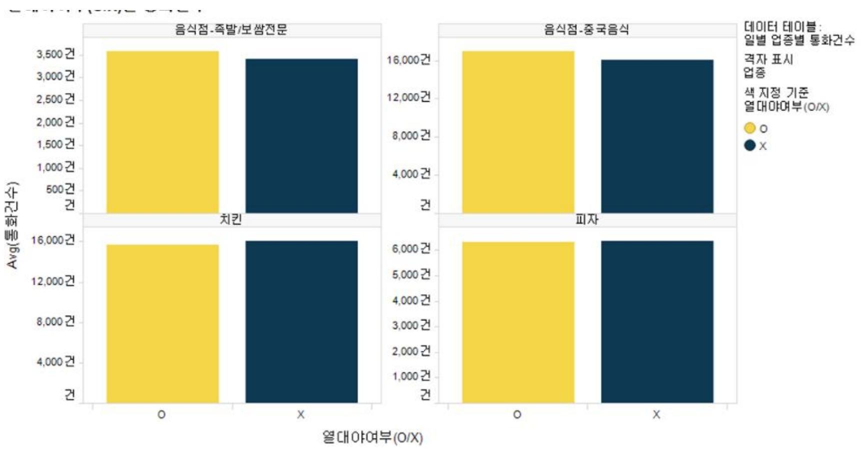


Figure 18
[Comparison of the call volume according to the occurrence or absence of tropical nights by industry]

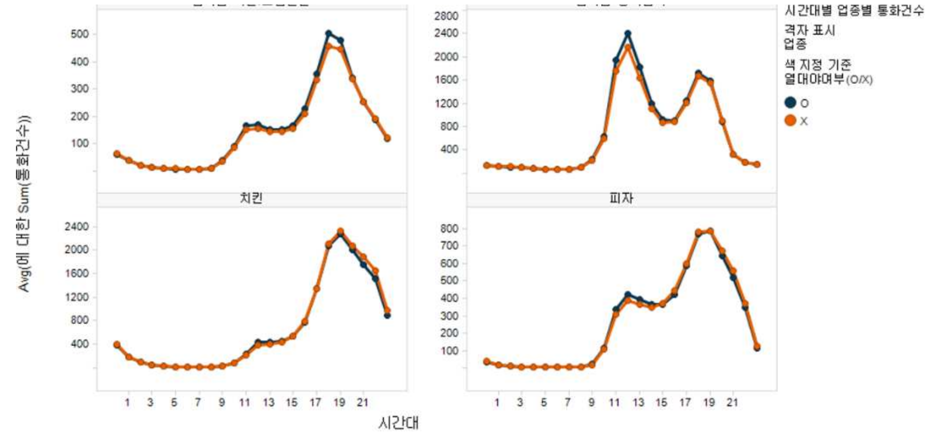


Figure 19 [Changes in call volume by time zone in case of tropical night]

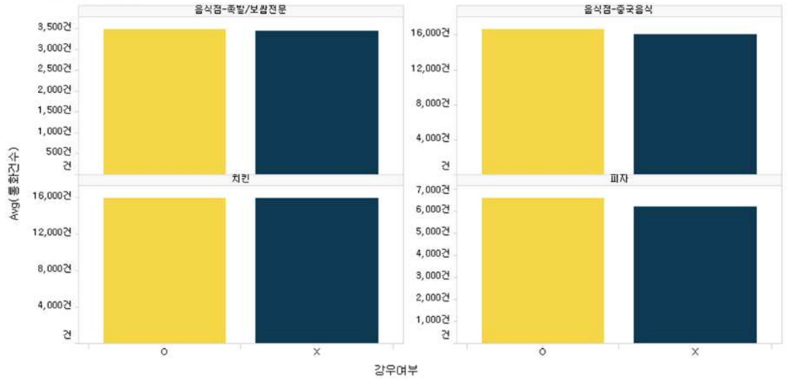


Figure 20 [Comparison of call volume according to rain or shine]

Although it was predicted that the number of currencies in each industry will be affected by the weather, it is difficult to say that there is a correlation with temperature due to low r2, and even in tropical nights and heavy rain, there is no big difference in the amount of calls.

2.3.9 Sports Event

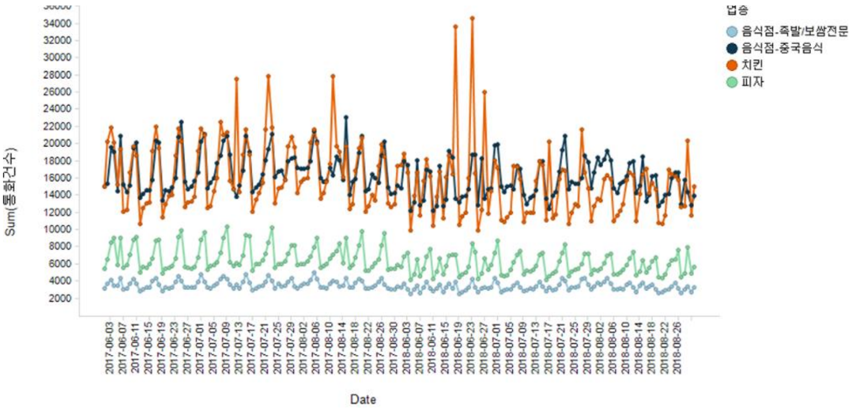


Figure 21 [Industry/Daily Currency Trend]

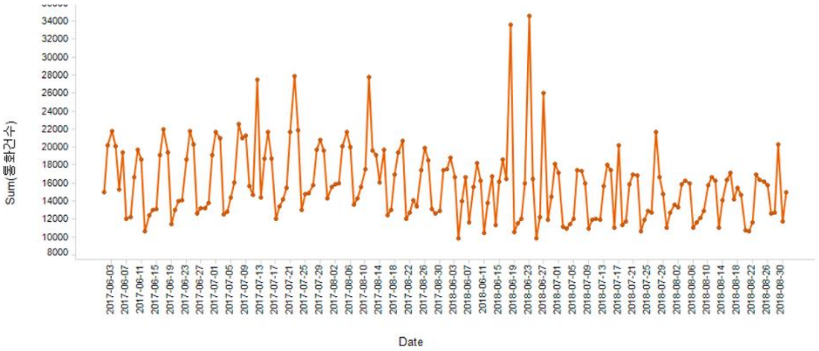


Figure 22 [Daily currency trends in the chicken industry]

Based on the line graph, the variation in the call volume is identified, and the pattern similar to that of the day-to-day call rate changes in the other three industries is repeated without any anomalies, but in the case of chicken, the call volume increases rapidly on a particular day.

In 2018, the number of calls for chicken was identified as the day when the Korean national team had a World Cup or Asian Games match, but in 2017, the number of calls surged even though there were no significant sporting events. Therefore, further investigation has shown that the day was first of the three dog days, second of the three dog days and last of the three dog days.

Chapter 3. Conclusion

33 billion delivery market and the entire industry estimates that around 20 trillion won in 2018 2013 in five years in application delivery market share in this growth to grow about 9 times. (national of delivery, Yogiyo, Badaltong) the three applications for existing delivery are waging fierce competition, and experts recently. Coupang, Naver, and Kakao, market entry by announcing extreme competition, the it will happen in this market in the future.

Application delivery platform in the market is growing fast delivery company and franchise businesses and future market delivery platform, partnering. You must take advantage of the growth.

In addition, the relevant companies collect customers' purchase information to survive in this market and use effective strategy and into action

3.1 Make Strategies

This analysis sets up a hypothetical new franchise and uses some information from this analysis to give directions to what strategies they might want to use to get into the market.

The new franchise company has decided to launch a chicken brand named A.

First of all, the companies should either build their own delivery platforms or adapt themselves to the market's flow-through partnerships with existing platforms, as 2.3.4 analysis determines that growth in the delivery application market is the most likely cause of the drop of more than 10 percent on-year in 2018.

Based on the analysis in 2.3.5 and 2.3.6, consider including a menu for female customers with relatively high sales weight. Of course, buyers and users should consider different cases, but since women in their 40s have the greatest consumption power and women have a clear preference for chicken over other industries, they might consider including a menu aimed at female consumers.

The promotion should be focused on the weekend. On weekends when there are virtually the most purchases, promotion through price discounts should be carried out to absorb the soaring demand at the weekend. In addition, when actual purchase decisions are made, media advertisements should be carried out during weekdays and weekends evening so that their products can be recalled only by internal searching without additional search processes.

In addition, if there are events such as three dog days in the chicken industry or sports events with high public interest, such as the national team's games, as the result of 2.3.9, the promotion should be carried out in accordance with these events, as they will have up to four times the sales increase compared to the existing sales.

Considering the location of the head office in the early stage of the project, 2.3.7 indicates that chicken calls have increased rapidly since 17:00 in particular the company is required to deploy direct branches in the top 10 areas where the volume of calls was high during the night.

3.2 Limitations

The analysis results do not reflect the characteristics of order-prone groups using the application, as data from the delivery application provider is not available during data collection, and the difference in characteristics that can occur depending on the carrier is unknown. In addition, the limitations of the collected data in analyzing multidimensional problems that take into account multiple variables did not result in quality results.

More quality data, such as delivery application data and carrier-specific order call data, can be obtained by conducting an analysis with additional consideration of variables that may affect delivery, which will result in a more reliable, high-quality analysis.

Reference Literature

- 고민경(2011). 배달음식의 이용실태와 영양정보표시 인식도
- 박준규(2012). 빅데이터를 위한 분석기술 활용방안 연구(석사학위). 세종대학교 대학원, 서울.
- 이희섭(2014). 빅데이터를 이용한 가뭄발생지역과 가뭄심도 분석(석사학위). 한서대학교 대학원, 서울.
- 통계청, 도소매업조사(2008), 서비스업조사(2017)
- 편집부(2012).『빅데이터와 DBMS 의 시장 전망』, 하연

Outlier Analysis of Nonferrous Data from Smart Manufacturing Plants

Sokchomrern Ean, Je-Seog Myeong, Kwan-Hee Yoo*

Dept. Of Computer Science, Chungbuk National Univeristy, South Korea
{chomrern, jeseogmyeong, khyoo}@chungbuk.ac.kr
*Corresponding Author

Abstract. The data that we use are viewed as one of the precious assets in the modern smart manufacturing system. The data can be formulated into useful information for visualizing and reporting for decision maker to make the right decision. However, from time to time the data can be violated by noise or error in the PLC (Programmable Logic Controller) or sensor devices. Therefore, this paper presents the methods of finding the most significant different data points in the set of data and is followed by illustrating the statistical visualization using box plot. Similarly, the paper highlights the experimental results with the attention to get high efficiency of productivity.

Keywords: Outlier Analysis, Nonferrous Data, Smart Manufacturing Plants

1 Introduction

The rapid growth of high tech in industrial production is emerged dramatically, many firms try to achieve the three main pillar goals which are quantity, quality and cost [1]. With this in mind, we propose to use the outlier analysis method to find out the most significant different incoming data from the machine in nonferrous factory. In the sense that the results will help the data operator and administrator to monitor some issues which are taking place on the shop floor. That said, getting the benefit of applying outlier analysis is essential to showcase and to maintain the high efficiency and sustainability in the plants. Ultimately, the main aim of the paper is to conduct the research for the sake of the following contributions:

- We select an outlier approach which is viewed as one of the most popular statistical approaches to show the crucial different of data [2].
- We examine on the findings to point out how much we achieve the goal of making the statistical visualization in the smart manufacturing firm.

After that, the paper is organized into five sections. Section 2 gives the overview of the related study. Section 3 describes the proposed method. Section 4 presents the experimental results of analyzing the data using outlier approach. Section 5 concludes the paper and takes the future work into account.

2 Related Study

In this part, the outlier mechanism has been considered as the tool to detect the uncommon data values in a specific period. To do so, there are many researchers have established and proved their algorithms to be most valid into the universal proposed methods. In [3], the author indicates that when some entities are far different from the other entities within the same group of data. In that case we can have a simple example as we imagine that there are five points that can be seen locating outside the boundary of the same zone of multiple data points. Those five points are considered as outliers.

So far, what we know about outlier is largely based upon empirical studies that investigate how to define the outlier in general. Apart from that, there is a large volume of published research studies describing the type of the outlier. In [4], the author proves that the outlier can be categorized into three kinds such as point, contextual and collective anomalies. First, the straightforward example is figuring out the fake of bank credit transaction which the expenditure amount is used as the main data. If the data appears in the inappropriate manner which can be defined as an outlier by taking the amount of spending is completely different from the rest of daily transaction [5]. Second, to better understand what the context is, we start looking at the case study of investigating the spending habits of the card owner. It is common that the amount of money which is spent in a period of holiday is higher than the daily's. However, it might be vague if the same amount of holiday expenditure keeps the same even during the working day [5]. Lastly, the collective anomaly outlier can be viewed as a group of data entities which gives a help in finding the significant differences of data [6].

3 Proposed Method

The proposed method of this study is used to analyze the outlier for monitoring the data in the nonferrous industry area. With the purpose of improving and sustaining the high efficiency of consuming electrical power in the production, the statistical method and data visualization jump into play as the effective indicator to assist the operator as well as the decision maker to obtain all the critical information of data which are inconsistent. To overcome the unexpected data retrieved, we apply the outlier method to the data coming from PLC of each machine in the flow of production lines.

One of the straightforward and powerful of statistical methods is box plot. Fig. 1 shows the nature of box plots and will be followed by the detailed explanation. Notably, there are five major numbers are needed to be paid much attention in the conceptual and practical definition. One is minimum number which can be seen at the bottom of the box. Next is the first quartile or twenty-five percent of the total. After that, the line which crosses the closed box is called median. The fourth number is named as the third quartile or it is equal to 75 percent of the whole numbers. Lastly, the maximum number is accounted for 100 percent. All things considered, if the data points are either greater than the maximum point or lower than the minimum point will be falling into the outlier category.

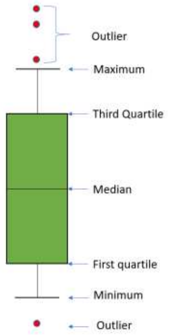


Fig. 1. Outliers appear in the boxplot diagram

4 Experimental Results

In this part, the finding of the paper draws the attention to the illustration of the outliers in the form of a box plot diagram. That said, the input features are playing a vital role to get insight into the procedure of detecting outlier in the set of data. Regarding to the dataset, the data are collected from the nonferrous factory which located in South Korea. The amount of data is accounted for 7,238,012 records. In addition, to generate the outcome, we use a desktop computer with specification such as CPU Intel® Core™ i7-6700 3.40GHz, RAM 16 GB, Graphics card NVIDIA GeForce 9800 GT, Operating System Windows 10 and Integrated Development Environment Jupyter notebook.

Table 1 shows the useful topics for illustrating the outlier. The first column is a topic. The OD is an abbreviation of outlier detection. The second column is the definition. Notably, the PPPD stands for power per product quantity. The last column indicates the figure of each topic.

Table 1. Outlier Detection Topic and Definition

Topic	Meaning	Figure
OD1	Outlier detection of PPPQ by product	Fig. 2
OD2	Outlier detection of product quantity by product and tap position	Fig. 3
OD3	Outlier detection of element power by product and tap position	Fig. 4
OD4	Outlier detection of mixed materials by PPPQ	Fig. 5
OD5	Outlier detection of PPPQ per tap position	Fig. 6

From Fig.2 to Fig. 6, it can be clearly seen that the outliers are represented by the black circles which are above the maximum point or lower the minimum point. Reference to these visualizations, it is useful to find out the causes which lead to inconsistent in the large data set. It may be caused by the fault of the machine or the error of manufacturer. For example, Fig. 2 shows that the product A consumes much power per product quantity. Moreover, if we compare the product A to the product C, there are considerable outliers on the product C even it consumes less power per product quantity. The following graphs also reveal the outliers with the specific tap position, so that we can monitor the machine effectively and efficiently.

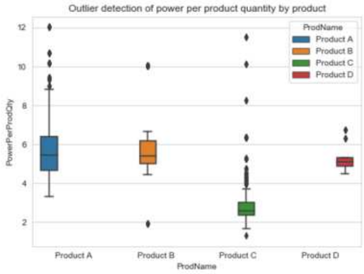


Fig. 2. Outlier detection of power per product quantity by product

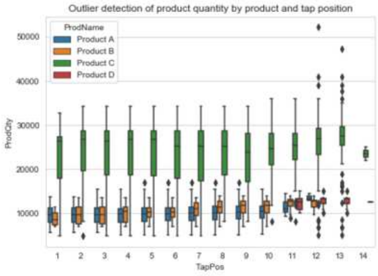


Fig. 3. Outlier detection of product quantity by product and tap position

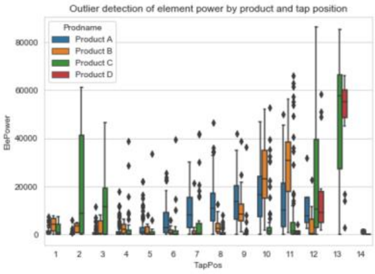


Fig. 4. Outlier detection of element power by product and tap position

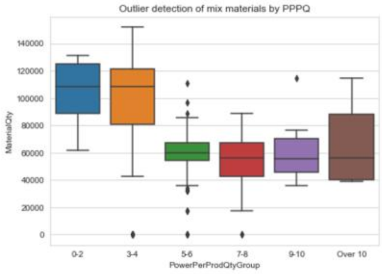


Fig. 5. Outlier detection of mixed materials by PPPQ

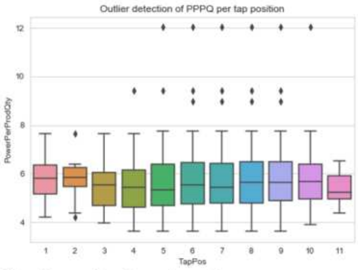


Fig. 6. Outlier detection of PPPQ per tap position

5 Conclusion

In conclusion, the paper shows how to get the benefit from applying outlier analysis to the smart factory system by associating with statistical methods and box plot visualization. Then again, the finding focuses on differentiating from normal data and anomaly data. Knowing that variability of data can serve as the tool to maintain the stability of process with less expense in terms of data consistency and data validation. The best way to separate the data can lead to how well the operation and management of the machine and management perform. Eventually, the further studies will be needed to consider with the sake of making a better smart system service.

Acknowledgments. This research was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program, No.10082578, 'Development of intelligent operation system based on big data for production process efficiency and quality optimization in non-ferrous metal industry'.

References

1. Ebrahimi, M., Armand, B., Eva, R.: A Roadmap for Evolution of Existing Production System Toward The Factory of The Future: A case study in automotive industry." In 2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), pp. 274-281. IEEE (2018)
2. Aggarwal, C.C.: Outlier analysis. In Data mining, pp. 237-263. Springer, Cham (2015)
3. Thennadil, S. N., Dewar, M., Herdsman, C., Nordon, A., Becker, E.: Automated weighted outlier detection technique for multivariate data. Control Engineering Practice, 70, pp. 40-49 (2018)
4. Aha, D. W., Richard, L. B: Feature Selection for Case-based Classification of Cloud Types. An empirical comparison. In Proceedings of the AAAI-94 workshop on Case-Based Reasoning, 106, pp. 112 (1994)
5. Choudhary, P.: Introduction to Anomaly Detection. Oracle DataScience.com, https://www.datascience.com/blog/python-anomaly-detection
6. Hoppenstedt, B., Reichert, M., Kammerer, K., Spiliopoulou, M., Pryss, R.: Towards a Hierarchical Approach for Outlier Detection in Industrial Production Settings (2019)

Deep Learning for Forecasting Production Quantity of Ferro-alloy in Electric Arc Furnaces (EAF)

O SangWon[1], Young Seog Yoon[2]*, Kwangroh Park[2]

[1] Electronics & Computer Engineering Department
CHONNAM NATIONAL UNIVERSITY
Gwangju, South Korea
[2] Future & Basic Technology Research Division
Electronics and Telecommunications Research Institute (ETRI),
218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, KOREA

osw0778@gmail.com,{isay, krpark}@etri.re.kr
*Corresponding Author

Abstract. Large fluctuations in production quantity of Ferro-alloy manufactured in an electric arc furnace (EAF) prevents manufactures to establish a reliable manufacturing and operating plan. However, little is known about the factors and their effects on Ferro-alloy production in EAF. Moreover, the nature of a furnace (extremely high temperature in a fully-closed space) reinforces the uncertainty. To fill out the gap, this study applied deep learning models to forecasting production quantity of Ferro-alloy. Specifically, we employed recursive neural network (RNN), Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) with two-year data of daily input power, daily pressing depth of three electrodes, the number of daily productions, and a dummy variable of consistency in producing same product. Prediction performance of deep learning models indicate that deep learning might provide a useful guideline for establishing manufacturing plan.

Keywords: electric arc furnace, Ferro-alloy, RNN, LSTM, GRU, forecasting, production quantity

1 Introduction

Although deep learning with big data is expected to provide the advanced analytics in a manufacturing landscape, lack of data and practical evidence obstruct small medium enterprises (SMEs) to adopt it. Moreover, previous studies in manufacturing industry mainly apply deep learning to defect classification [1], energy saving [2], and image detection [3]. Deep learning is rarely adopted in forecasting production quantity because it depends on traceable and measurable input such as raw material, labor, cost, and investment.

However, for Ferro-alloy manufacture's perspective, it is one of the significant challenges to forecast production quantity accurately because of three reasons; first, it is infeasible or complicated to measure input variables related to products in fully closed electric arc furnaces (EAF). For example, the temperature in electric furnace reaches 2,000 ºC, so that it is almost impossible to monitor and measure the internal situation in EFA (e.g. the exact position of electrodes). Second, little is known about the factors influencing on production quantity of Ferro-alloy. Only few factors are analyzed yet [4-6], and most of factors are hardly measurable. Third, the variations range of production quantity of Ferro-alloy is immensely huge. As seen in figure 1, the wide fluctuations are observed.

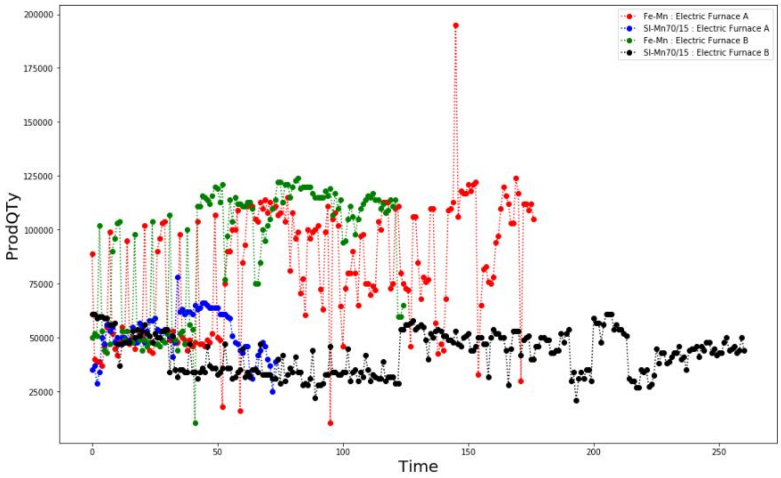


Fig. 1. Fluctuation in product quantity of Ferro-alloy (source: data used for this study)

Because of large fluctuations, it is fundamentally required for the manufactures to forecast the production quantity to establish a reliable production plan, to operate furnaces efficiently, and to allocate resources appropriately. To fill out the gap, this study aims at forecasting production quantity of Ferro-alloy (e.g. Fe-Mn and Si-Mn) by applying deep learning models for prediction.

2 Literature Review

2.1 Characteristics of Electric Arc Furnace (EAF)

Electric Arc Furnace (EAF) is widely used to melt multiple ore stones in order to improve physical characteristics of them or to eliminate impurities by the exothermic reaction generated by high-voltage electricity. In order melt multiple type of ore, EAF provides high-voltage electricity on electrodes. For the case of Ferro-alloy (i.e. Fe-Mn, Si-Mn), the company produces the products three to five times a day. Ore stones are continuously supplied. 3-phase AC power is supplied, and then the arc occurs when electrodes approach coke. The generated arc melts ore stones in EAF. After finishing the melting process, the melted stones is discharged into mold box through tap hole.

After natural cooling process, the melted stones is converted into products. The structure of EAF is depicted in figure 2.

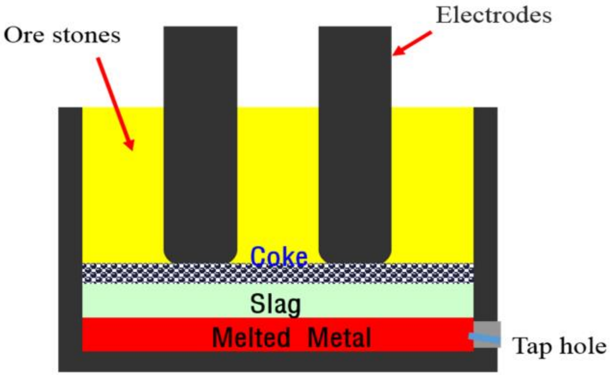


Fig. 2. Structure of EAF

Because there are great variations in amount of production quantity in a factory, it is fundamentally required for manufactures to forecast the production quantity accurately. The factors influencing on production quantity has not yet fully studied. Previous studies mainly focus on reducing slag for protecting environment [7, 8] rather than those factors. Only few studies were conducted for those factors except the transferred power and the nearness of electrodes on coke [4, 6]. However, other factors are still uncovered. Because of the extremely high temperature and the entirely closed structure of EAF, it is almost impossible to monitor the inside of EAF when in processing.

2.2 Recursive Neural Network (RNN)

Current study applied deep learning models to forecasting production quantity of Ferro-alloy. Recursive Neural Network (RNN) is one of the most widely used learning models. Because of its ability of accurate prediction and classification[9], especially for sequential events, RNN has been widely applied into various fields[10]. RNN processes information sequentially and reflects previous computation results. Hidden state in RNN retains information from previous calculation and it is used the current input.

One of the most powerful advantages of RNN is flexibility and allowance of variety in modellings because of its simple structure of entering input and output regardless of the length of sequential data. Accordingly, RNN is suitable for learning sequential data.

2.3 Long Short Term Memory (LSTM)

Although RNN is appropriate for learning sequential data over time, the memorized information can't be reflected efficiently if data size is large or the extended time interval between earlier layer and current layer is far. The vanishing gradient problem becomes worse as the number of layers increases.

In order to resolve this problem, LSTM was proposed [11] with the idea of forgetting unnecessary information for efficient calculation. Accordingly, learning performance of LSTM is reliable and efficient while it is originated from RNN architecture. LSTM have been widely adopted for learning time series data.

2.4 Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) adopts the main idea of LSTM, but its structure is more simpler than LSTM to achieve calculation efficiency [12]. Because of the simplified structure, GRU is known to improve accuracy and efficiency of learning. Thus, it has been widely used in researches, especially for analyzing large size of data.

3 Data description

3.1 Data source, collection, structure

This study aims at forecasting product quantity of Ferro-alloy manufactured in electric arc furnaces by employing RNN, LSTM, GRU deep learning models with time-series data consisting of several input. Two-year data was provided by one of the largest manufactures in Korea. The company records input raw material, output (production quantity), and relevant information for operating two factories. The manufacture produces various Ferro-alloy (i.e. Fe-Mn, Si-Mn 70/15, Si-Mn 70/14, and Si-Mn 60/14.)

The number of productions varies because production cycle is not static. Specifically, the number of productions ranges from one to five in a day. Factory workers manually record the production quantity per each production cycle to manage furnace factory operation. The daily report contains the following data of:

1) *Factory code*
2) *Manufacturing product*
3) *Date of production*
4) *Number of daily productions*
5) *Start and end time per each production*
6) *Production quantities per each production*
7) *Input power*
8) *Daily intrusion depth of three electrodes*
9) *Input ore stones (quantities)*

Based on the daily report, we constructed a database by converting manual recorded data into digitalized data.

3.1 Descriptive Statistics of Data

Although little is known about the factors influencing on production quantity of Ferro-alloy except findings from previous studies [4, 6], tacit knowledge of production allows us to determine input data for this study. We asked anonymous factory experts, and they pointed out the gap between electrodes and coke, input power, the number of productions, and consistency in producing same product might influence on production quantity regardless of final products.

Hence, we posit that production quantity is highly associated with 1) input power, 2) average of pressing depth of three electrodes, 3) number of productions, and 4) consistency in producing same product (if same product is manufactured in previous production: 1, otherwise: 0.)

Totally, 810 daily reports data recorded from Jan. 2017 to April in 2019were used as input data. Although each factory manufactures several Ferro-alloy metals, we only used data of two products (i.e. Fe-Mn and Si-Mn 70/15) because the numbers of other production seem to be not enough to apply deep learning. The number of Fe-Mn and Si-Mn 70/15 are 302 and 336, respectively. Table 1 summarized the descriptive statistics of data.

Table. 1. Descriptive statistics of data

				Min	**Max**	**Average**	**Std.**
Input	Fe-Mn	Factory A	Daily Input Power (IP)	62,600	328,200	226,335.5	64,209.3
			Avg. of pressing depth of three electrodes (DEP)	0	680	230.6	129.3
		Factory B	IP	121,400	316,700	243,852.8	66,369.6
			DEP	20	820	244.6	147.0
	Si-Mn70/15	Factory A	IP	182,400	300,900	252,024.0	29,432.5
			DEP	30	600	297.3	121.2
		Factory B	IP	121,800	295,000	220,359.0	40,609.6

			DEP	30	670	306.9	122.9
	Number of daily production			1	5	3.6097	0.7695
	Dummy variable (consistency in producing same product)			If same: 1, otherwise: 0		0.9655	0.1825
Out put	Fe-Mn	Factory A	Production Quantity (Q)	10,500	195,000	82,223.2	29,043.1
		Factory B	Q	10,700	124,000	92,405.6	29,169.7
	Si-Mn70/15	Factory A	Q	25,000	78,000	51,480.0	10,332.6
		Factory B	Q	21,000	61,000	42,611.1	9,488.1

4 Experiments

4.1 Hyper Parameters

In general, forecasting performance in deep learning models depend on hyper parameters of the number of hidden units, the number of stacked layers, the number of epochs, and learning rate [ref]. Except LSTM specific hyper parameters (forget bias, keeping probability), we set the same values for each parameter in order to avoid the effect of them on forecasting results. We have modified hyper parameters to figure out the values for optimization through trial-error approach. We conducted our experiments in Python 3.6 with TensorFlow 1.14.0. The values adopted for hyper parameters are summarized in table 2.

Table. 2. Hyper parameters and values

Hyper parameter	Value	Adopted in
number of hidden units	20	RNN, LTSM, GRU
number of stacked layers	4	RNN, LTSM, GRU
number of epochs	15,000	RNN, LTSM, GRU
learning rate	0.01	RNN, LTSM, GRU
forget bias	0.5	LSTM only
keeping probability	1.0	LSTM only

4.2 Experiment Results

We split our data into ratio of 7:3 for train and test data. Then, we conducted scaling data into [0, 1] for efficient computation and prevention of overflow and underflow. In order to evaluate forecasting accuracy for each model, we adopted three evaluation criteria of Root Mean Square Deviation (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). They are commonly adopted to measure the accuracy of deep learning results. The evaluation results for each model are summarized in table 3 to 5, respectively.

Table. 3. Prediction accuracy of RNN for two products

Product	**Factory**	**Data size**	**RMSE**	**MAE**	**MAPE**
Fe-Mn	A	177	27,528.47	18,339.67	22.56
	B	125	15,476.33	9,365.18	10.88
Si-Mn70/15	A	75	9,314.65	7,472.45	19.16
	B	261	6813.67	5055.65	13.92

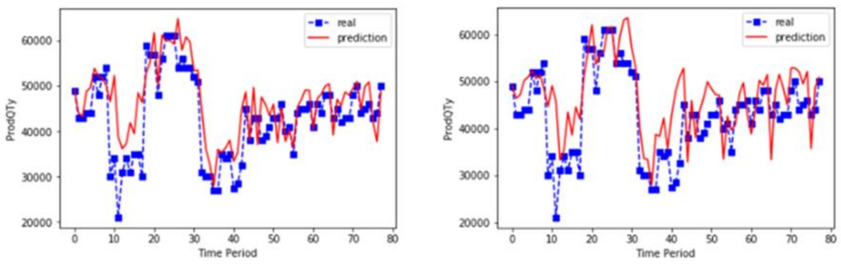
Table. 4 Prediction accuracy of LSTM for two products

Product	**Factory**	**Data size**	**RMSE**	**MAE**	**MAPE**
Fe-Mn	A	177	22,137.35	14,735.54	18.44
	B	125	16,564.90	10,202.09	11.63
Si-Mn70/15	A	75	11,121.77	9,347.67	21.58
	B	261	6,239.38	4,560.52	12.48

Table. 5 Prediction accuracy of GRU for two products

Product	**Factory**	**Data size**	**RMSE**	**MAE**	**MAPE**
Fe-Mn	A	177	26,437.78	16,806.76	20.43
	B	125	16,196.15	10,460.04	11.85
Si-Mn70/15	A	75	5,677.59	4,525.90	10.63
	B	261	7,353.39	5,625.01	15.44

As seen in table 3 to 5, prediction based on LSTM model shows the best performance in terms of forecasting accuracy. In addition, learning results for Si-Mn is more accurate than that for Fe-Mn regardless of factories. Figure 3 compares the predicted results of LSTM and GRU with the real production quantity of Si-Mn70/15.



< LSTM>   < GRU>

Fig. 3. Comparison prediction results (LSTM, GRU) with real production quantity of Si-Mn70/15 in Factory B

5. Conclusions

The use of big data and deep learning is expected to stimulate the innovation in manufacturing industries. In a case of Ferro-alloy manufacturing in EAF, it is essentially required to apply these techniques to forecast production quantity because of the extremely high temperature, entirely closed structure of EAF, wide fluctuations of production quantity, and limited knowledge on factors of producing Ferro-alloy in EAF. Therefore, manufactures suffer from uncertainty and unpredictability when they are establishing manufacturing plan.

To fill in the gaps, this study applied deep learning models to forecasting daily production quantity of Ferro-alloy. Even though there are great fluctuations on production quantity, it has not been fully investigated yet. We conducted RNN, LSTM, and GRU with data recorded in 638 daily report. We posit that daily production quantity is deeply associated with daily input power, daily pressing depth of three electrodes, the number of daily productions, and a dummy variable of manufacturing the same product consistently. Although forecasting results based on three models are acceptable, there might be unexplored factors influencing on production quantity. However, we showed that deep learning might be used to establish a basic guideline for scheduling production quantity under extremely uncertain conditions.

Although we provide meaningful contributions, there are several limitations. Further study could replicate our study with another input data since there exist very few related articles. Second, this study used small amount of data for training. Third, input data could be contaminated since the original data was written manually. Those limitations indicate that there are valuable research opportunities in a manufacturing industry.

Acknowledgments. This work was supported by Industrial Strategic Technology Development Program (10082578, Development of intelligent operation system based on big data for production process efficiency and quality optimization in non-ferrous metal industry) funded by the Ministry of Trade, Industry & Energy)

References

1. Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.: Traffic Flow Prediction With Big Data: A Deep Learning Approach. IEEE Transactions on Intelligent Transportation Systems 16, 865-873 (2015)
2. López, K.L., Gagné, C., Gardner, M.: Demand-Side Management Using Deep Learning for Smart Charging of Electric Vehicles. IEEE Transactions on Smart Grid 10, 2683-2691 (2019)
3. Zhou, S.K., Greenspan, H., Shen, D.: Deep learning for medical image analysis. Academic Press (2017)
4. Khoshkhoo, H., Sadeghi, S.H.H., Moini, R., Talebi, H.A.: An efficient power control scheme for electric arc furnaces using online estimation of flexible cable inductance. Computers & Mathematics with Applications 62, 4391-4401 (2011)
5. Hauksdóttir, A.S.a., Gestsson, A., Vésteinsson, A.: Current control of a three-phase submerged arc ferrosilicon furnace. Control Engineering Practice 10, 457-463 (2002)
6. Samet, H., Ghanbari, T., Ghaisari, J.: Maximizing the transferred power to electric arc furnace for having maximum production. Energy 72, 752-759 (2014)
7. Schoukens, E., AFS, S.: The Enviroplas process for the treatment of steel-plant dusts. Journal of the Southern African Institute of Mining and Metallurgy 93, 1-7 (1993)
8. Branca, T., Colla, V., Valentini, R.: A way to reduce environmental impact of ladle furnace slag. Ironmaking & Steelmaking 36, 597-602 (2009)
9. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Cognitive modeling 5, 1 (1988)
10. Hüsken, M., Stagge, P.: Recurrent neural networks for time series classification. Neurocomputing 50, 223-235 (2003)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9, 1735-1780 (1997)
12. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)

A study on operation regression model in electric furnace using data of production

Seonjong Bong[1] , Suyoung Chi[*]

[1] Korea University of Science and Technology, 21 7,Gajeong-ro,Youseong-gu, Daejeon, Korea,
(Tel : +82-010-6394-4469;E-mail: sunjong@ust.ac.kr)
[*] Electronics and Telecommunications Research Institute, Dajeon,305-370, Korea
(Tel : +82-02-042-860-5337;E-mail: chisy@etri.re.kr) *Corresponding author

Abstract. One of the most important issues in the 4th industry is Big Data Analysis. generally, we have increased the production efficiency through the know-how of the workers, now research is underway to find the optimal process operation by analyzing big data. There are a lot of companies that want to store and analyze many data coming from sensors and make models and apply them to their operation. We analyzed the data by using a general linear regression model estimation method and we propose the new model to make regression model. However, in a real environment it is difficult to find a linear regression model that fully satisfies these data because of human factors, data anomalies, and factors that can result from inaccurate data, depending on the choice of data scientists to analyze the data have. In this paper, we compare with ordinary model using production data.

Keywords: Data analysis, Regression model, Big data

1 Introduction

One of the most important issues in the 4th industry is Big Data Analysis. So far, if we have increased the production efficiency through the know-how of the operators, now research is underway to find the optimal process operation by analyzing big data. There are a lot of companies that want to store and analyze many data coming from sensors and make models and apply them to the field. Meanwhile, the steel industry uses electricity to operate, and electric resistance companies are demanding operation models that reduce electricity consumption while increasing production. However, studying the operation of a furnace has many limitations. Among them, it is difficult to attach the sensor due to high temperature due to electric resistance, and it is most difficult that the inside of the resistance furnace is unknown because the inside is filled with raw materials. Therefore, we collect data by strapping various sensors and indirectly understand the production process.

In this paper, we try to check the relationship between electricity usage and production before modeling the operating process with data collected from various sensors. In particular, we try to compare the results of various linear regression algorithms with one another to find a linear regression algorithm that can correct for errors that can occur in a large number of data. We can obtain good results by modeling the operation process according to the analysis result.

2 Preprocessing data

Using the logbook provided by the company, data on electricity usage and production for three months were obtained. It is divided into 300 train data and 50 test data. 300 train data have a Pearson correlation coefficient of 0.6.

- Pearson's correlation coefficient : 0.6087

Pearson's correlation coefficient is a measure of the correlation of two independent variables by standardizing the covariance. If the numerical value is more than 0.6, it means that it has some linear relationship as shown in the following table.

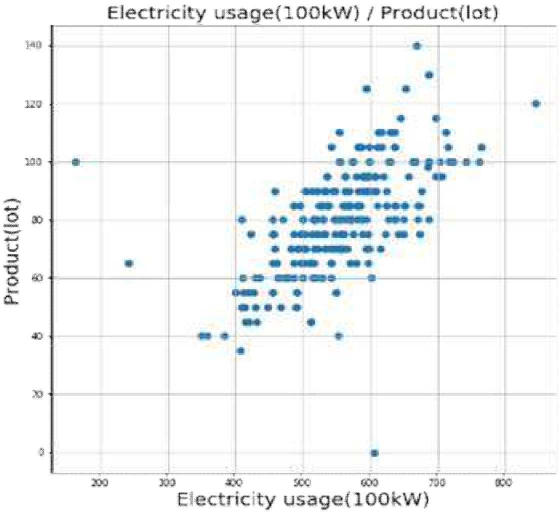


Fig. 1. Electricity usage & Product data

Linear regression is a regression technique that models the linear correlation between dependent variable Y and one or more independent variables X. Train data were used to construct a model using well-known linear regression model estimation methods and experiments were conducted to determine the accuracy using test data.

3 Experiment

Linear regression is a regression technique that models the linear correlation between dependent variable Y and one or more independent variables X. Train data were used to construct a model using well-known linear regression model estimation methods and experiments were conducted to determine the accuracy using test data.

3.1 Mean linear regression model

In general, the more data there is, the more the model converges. Especially in the case of big data, it is higher than the data which is less likely to converge into one model. Using this, we first made a model with total electricity usage and total output.

- Total electricity usage : 166250(kW)
- Total output : 24398(lot)

$$y = 0.14675x \qquad (\mathbf{1})$$

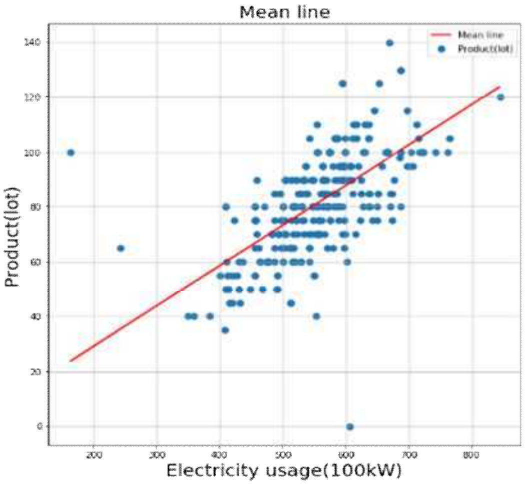


Fig. 2. Mean linear regression model

Using the above equation, we get the following results. If we visualize the data, we can get the following results.

3.2 Least squares & Robust method regression

Second, the linear regression model was estimated using Least squares method regression. The slope of the equation is as follows, and the y intercept has the following values.

$$y = 0.13537x + 6.30577 \qquad (2)$$

$$y = 0.15312x - 3.49537 \qquad (3)$$

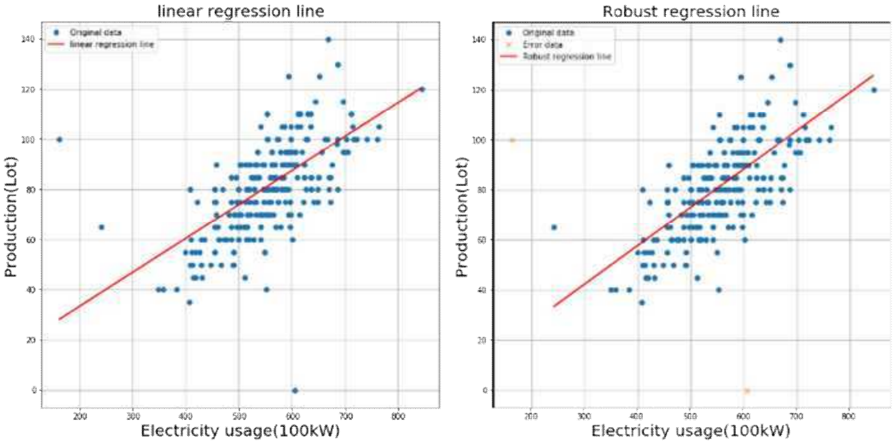


Fig. 3. Least-squares reg model(left) & robust reg model(right)

The model is as follows. Generally, least squares method regression has a disadvantage, and Robust regression is a model made to complement it.

3.3 Count_max Regression

To make the fourth model, we made one assumption. The regression curve that reflects all the data is, in the end, approximate, indicating the direction of the test data. We plotted a linear regression curve that traverses the most points if we draw a one-dimensional regression curve with Train data. It should be noted, however, that the greatest number of slopes are 0 or -1, which is the most common. The other linear regression models above pass through four or six points through the train data, while our linear regression model passes through the fifteen points. The equation is as follows.

$$y = 0.36210x - 111.53243 \qquad (4)$$

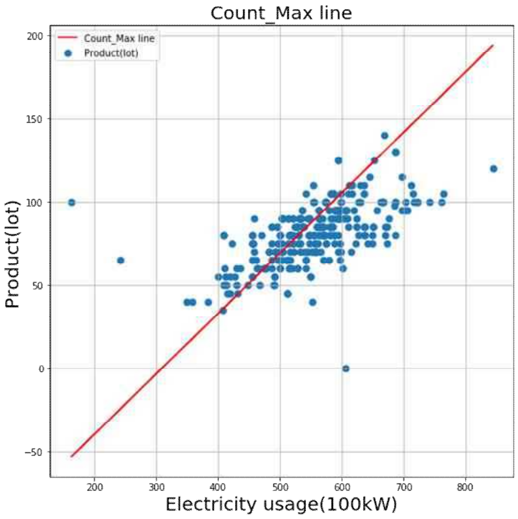


Fig. 4. Count-max regression model

3.4 Experiment analysis result

We compared the yields with the test data by estimating the linear regression model made by the above four models. The average difference in the production of the fourth model we propose resulted in worse results than in the other models, but when we plotted the results of the test, the number of production matches was 5 times higher than the other models.

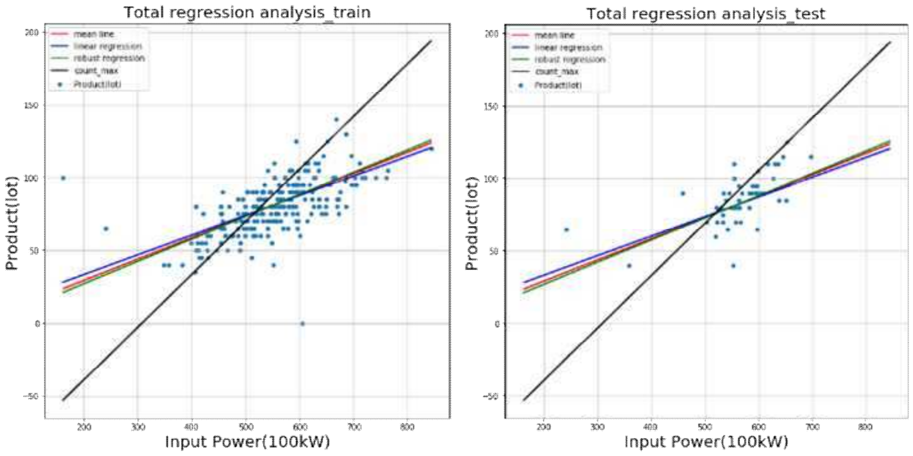


Fig. 5. Regression model(Train & Test data)

Table 1. Product Accuracy

Model	Product Accuracy	
	Mean of Product differ	***Accordance of test data***
Mean	-3.46252(lot)	10%(3/30)
Linear	-3.57191(lot)	10%(3/30)
Robust	-3.36643(lot)	10%(3/30)
Count_max	6.41343(lot)	17%(5/30)

4 Conclusion

We analyzed the data by using a general linear regression model estimation method. We will try to further develop the new Count_max model we have assumed. In a real environment, however, it is difficult to find a linear regression model that fully satisfies these data because of human factors, data anomalies, and factors that can result from inaccurate data, depending on the choice of data scientists to analyze the data have. In the case of Big Data, these parts are more important, and depending on which model you choose, the accuracy of the results will vary.

Acknowledgments. This material is based upon work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program. No. 10082578, 'Development of intelligent operation system based on big data for production process efficiency and quality optimization in non-ferrous metal industry.

References

1. Sivri, M. S., & Oztaysi, B. Data analytics in manufacturing. In Industry 4.0: Managing The Digital Transformation (pp. 155-172). Springer, Cham (2018)
2. Suykens, J. A., & Vandewalle, J. Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300 (1999)
3. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association, 74(368), 829-836 (1979)
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830 (2011)
5. Waller, M. A., & Fawcett, S. E. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), 77-84 (2013)

Working Pattern Analysis of Nonferrous System Using Statistical Analysis

Sokchomrern Ean, Je-Seog Myeong, Kwan-Hee Yoo*

Dept. Of Computer Science, Chungbuk National Univeristy, South Korea
{chomrern, jeseogmyeong, khyoo}@chungbuk.ac.kr
*Corresponding Author

Abstract. Using the statistical analysis method in developing the nonferrous system is crucial for maintaining the stability and improving the efficiency of production products with the highest effective energy consumption. That said, the process of doing that is quite complicated and requires a lot of diverse expertise and skills. In this paper, we give an overview of finding standard working pattern analysis of tap position and skip time from the large dataset of nonferrous environment.

Keywords: Working Pattern Analysis, Nonferrous System, Statistical Analysis

1 Introduction

In the history of system management development, a working pattern has been thought of as a key factor in establishing the reliable platform and framework for many sectors including private and public organizations [1]. Furthermore, recently researchers have examined the effect of working pattern on the productivity along with defining the best model in practice to reduce the high cost of making the product. Taken together, the nonferrous firms have become an interesting field to observe the pattern of working. Importantly, being aware of the movement of tap position from one tap to another is becoming more common to take control of the recovery rate effectively [2]. Likewise, study the skip time between the tap positions is not only good for the production, but also better to know the information of electricity consumption to achieve the minimization of energy in operating the machines.

In this paper, we aim to investigate the pattern of tap position and skip time with the attention to increase the productivity and reduce the energy consumption. Having said that, the working pattern can tell how well the development process is conducted to achieve the final goal of improving the productivity and cutting the cost of production. Next, the outstanding contributions of this paper are shown as below:

- We start by proposing an approach to visualize the tap position in a form of chart with the useful information, including tap position number, frequency of tap position, stay time, element power and is followed by the average of each feature.
- We examine the pattern of tap position with respect to the specific product name and within the period of each month.

- We get into the method of determining the standard pattern of tap position as well as the skip time of tap position.

Next, the paper is divided into five sections. Section 2 describes the overview of the related study. Section 3 highlights the proposed method. Section 4 shows the experimental results of working pattern. Section 5 summarizes the paper and further suggests the future work.

2 Related Study

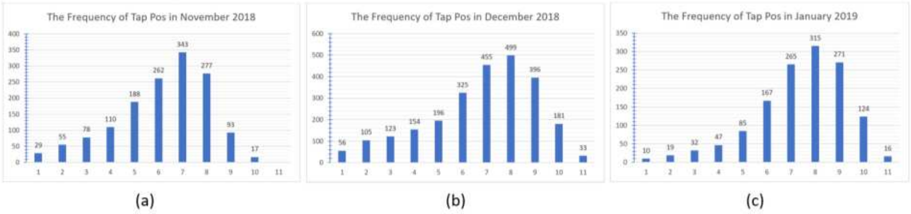
Developing a standard pattern, we first need to understand the outstanding trend of input, process and output. Apart from that, the previous studies are likely interested in process monitoring rather than looking into the general pattern. In [3,4], the authors suggest that the statistical process monitoring should be introduced to find out the fault in the process with the help of statistical pattern analysis and big data technology. By doing so, the factory will gain a lot of benefits, especially smart manufacturing monitoring development and online data monitoring.

Another previous research, the author tries to compare the principal component analysis (PCA) and the statistics pattern analysis (SPA) in terms of monitoring the process variables to implement the pattern analysis of fault detection in smart monitoring system. The experimental results prove that the SPA is better than PCA if we perform the nonlinear dynamic process[5].

Overall, the main purpose of developing the standard pattern in the smart manufacturing system is to find the best way to work in the production with the attention to save time and cost. Significantly, the main aim is to make use of the energy consumption effectively and efficiently. By doing so, the quality of work and output will move to the advance development of modern working pattern with the cutting-edge technology and innovation with the combination of the new process design [6]

3 Proposed Method

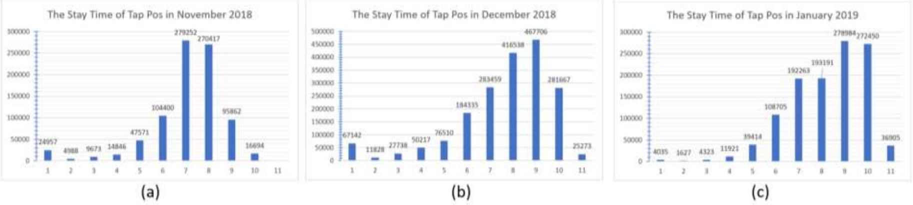
The proposed of this paper is used to determine the working pattern of the tap position movement and the skip time. In fact, the knowledge of tap position is crucial to get an understanding of the pattern. Regarding to this, we start looking closer to the frequency of each tap position within a period of each month. To give an illustration of what the charts mean in Fig. 1, we first look at the x-axis and y-axis. The x-axis is used to denote the tap position number while the y-axis represents the frequency of tap position. Firstly, we select the data on tap node from November 2018 to January 2019. After that, we filter out the product, namely Product A for pointing out the trend. Fig. 1 shows that there has been a gradual increase in the frequency of tap position from the tap 1 to the tap 7 and 8. After that, the number of tap position is projected to decline steadily.



(a) (b) (c)

Fig. 1. Frequency for each tap position in a working pattern

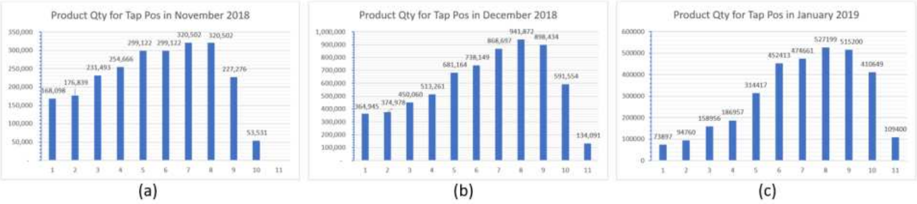
Fig. 2 depicts the charts of stay time of tap position. Units are measured in seconds. What is interesting is that if we compare the charts between Fig. 1 and Fig. 2, the frequency of tap position alone is not the causative factor of the stay time increases. For instance, in Fig. 1 (b), the highest frequency of tap position belongs to tap 8, but the longest stay time is the tap 9 in Fig. 2 (b).



(a) (b) (c)

Fig. 2. Stay time for each tap position in a working pattern

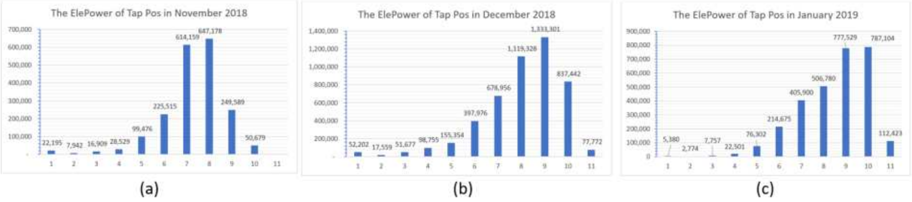
Fig. 3 gives the information about the product quantity at various tap positions. The charts present that there has been a marked rise in the quantity of product from tap 1 to tap 8. After that, the product quantity is likely to drop steadily.



(a) (b) (c)

Fig. 3. Product Quantity for each tap position in a working pattern

Next, Fig. 4 shows the element power which is consumed by tap position. The unit is measured in kilowatt per hour. A total of element power from tap 1 until tap 5 is slightly lower if we compare to the total of element power which is consumed by tap 7,8,9 and 10. Eventually, the energy consumption drops dramatically in the final tap.



(a) (b) (c)

Fig. 4. Element power for each tap position in a working pattern

The power unit per product quantity is also paid much attention to be aware of the trend for each tap position. In general, the charts in Fig. 5 formulate that there is an upward trend from the tap 1 until the tap 8, and then the remained taps are relatively low in the power unit per product quantity.

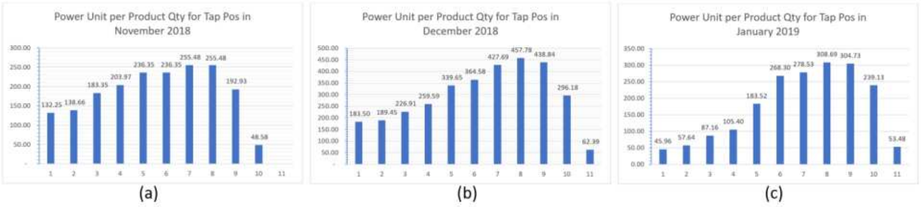


Fig. 5. Power unit per product quantity for each tap position in a working pattern

Next, it is decided that one of the best methods to adopt for this investigation is to point out the movement of tap position. Since the tap position is the leading cause of energy consumption in producing the specific product. In addition, criteria for knowing the movement of tap position are as follows: We denote i as the current tap position. Therefore, there are three possible directions which tap i can move to. The first scenario, the tab i moves to the next tap which is tap $i + 1$. The second scenario, the tap i can move to the previous tab node namely $i - 1$. The last scenario, the tap i moves to itself (see Fig. 6)

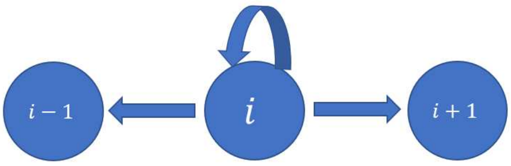


Fig. 6. Pathway of tap position in a working pattern

For the purpose of the pattern analysis, every tap position is illustrated as in the Fig. 7. The significance tap position which consumes considerable power and longer duration of stay time is depicted in a large circle using Arc Diagram. The path from one tap to neighboring tap is denoted by a curved line. From Fig. 7, it can be discovered that the tap 9 consumes much power and has very long stay time. It happens the same for tap 8 and 10.

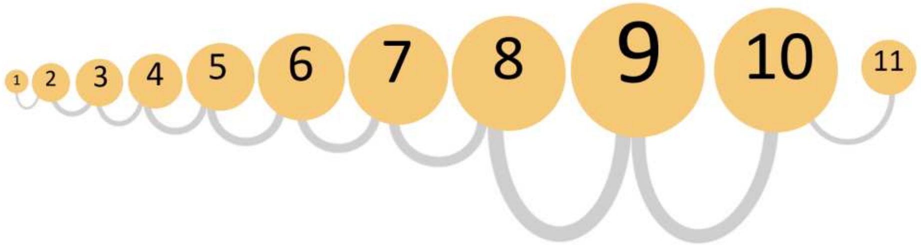


Fig. 7. Connection of tap positions in a working pattern

Following this, the forward and backward movements are given by the curved arrow vector as drawn in the Fig. 8. In this diagram, each tap position consists of four arrows. Two arrows are departing from the tap, and the other two are coming to the tap.

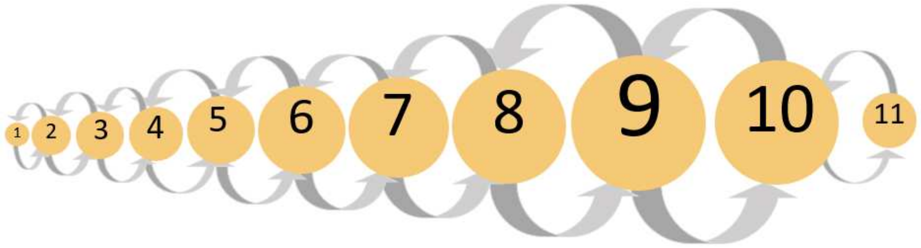


Fig. 8. Movement direction of tap positions in a working pattern

4 Experimental Results

In this part, the experimental results of the paper describe the five categories of tap position and the skip time. To conduct this research, the data are provided by the nonferrous metal industry. There are 8,137,023 observations from different tables in the same schema. Additionally, to perform the simulation, we use a computer with specifications such as CPU Intel® Core™ i7-6700 3.40GHz, RAM 32 GB, Graphics card NVIDIA GeForce GTX 750, Operating System Windows 10 and IDE Jupyter Notebook.

In this study, we divide the pattern into five types which are the worst, bad, so so, good and the best. The much element power and the longer stay time are, the most unwanted patterns are. In contrast, very less element power is, the better patterns we need to achieve. Overall, the worst pattern indicates that there are huge energy consumption and longest stay time while the following patterns represent less element power consumption and short of stay time. Thus, the most less power and a little stay time are considered the best pattern.

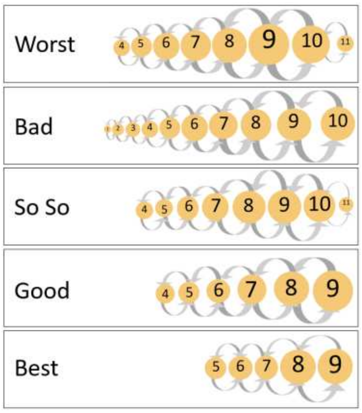
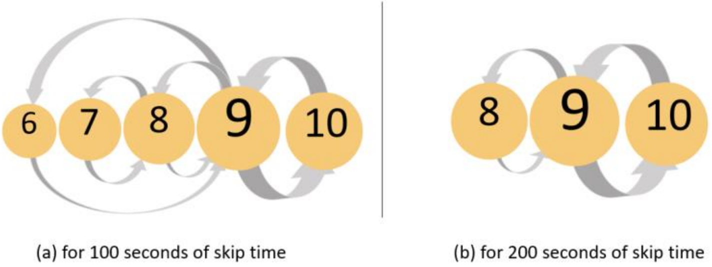


Fig. 9. Working patterns in five different levels

Although this kind of pattern formulation is considered normal in some cases, it seems reasonable to give the value of the skip time. For instance, if the skip time is set to 100 seconds, the tap that obtains stay time less than 100, it will not be shown in the diagram. The results of (a), as shown in Fig. 10, importantly indicate that the tap 9 can skip the previous taps and link to tap 6 directly. Another graph also has the same meaning, but in this case the skip time is 200 seconds (see Fig. 10 (b))



(a) for 100 seconds of skip time

(b) for 200 seconds of skip time

Fig. 10. Working patterns by giving the skip time

5 Conclusion

In summary, this paper sets out to determine the working pattern analysis in the sense that the possibility to reduce the cost of production in terms of minimizing the energy consumption. Moreover, to improve the productivity can be achieved by implementing the standard pattern with the help of finding the tap position movement and skip time. Taken together, these results suggest that using the diverse methods can be a better solution rather than focusing on the electricity power alone. However, these findings have some limitations in terms of interpreting the results. Therefore, further research is required to determine the working pattern.

Acknowledgments. This research was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program, No.10082578, 'Development of intelligent operation system based on big data for production process efficiency and quality optimization in non-ferrous metal industry'.

References

1. He, Q.P., Jin, W.: Statistical Process Monitoring as a Big Data Analytics Tool for Smart Manufacturing." Journal of Process Control, pp. 35-43 (2018)
2. Wang, Liqiang, Juncheng L., Anfa, L., Shenqiu, Z. Effect of Pouring Temperature on Microstructure and Mechanical Properties of Zr-Based Amorphous Alloys. In IOP Conference Series: Materials Science and Engineering, 394(3), pp. 032121. IOP Publishing (2018)
3. Galicia, H.J., Peter, H., Jin, W.: A Comprehensive Evaluation of Statistics Pattern Analysis based Process Monitoring." IFAC Proceedings, 45(15), pp. 39-44 (2012)

4. He, Q. P., Jin, W.: Statistics Pattern Analysis: A Statistical Process Monitoring Tool for Smart Manufacturing. Computer Aided Chemical Engineering, pp. 2071-2076. Elsevier (2018)
5. Wang, J. He, Q.P.: Multivariate statistical process monitoring based on statistics pattern analysis. Industrial & Engineering Chemistry Research, 49(17), pp.7858-7869 (2010).
6. Mabkhot, M., Abdulrahman, A., Bashir, S., Hisham, A.: Requirements of the smart factory system: a survey and perspective." Machines 6(2), pp. 23 (2018)

Development of Efficiency Improvement of Nonferrous Metal Production Process based on Big Data Analysis.

Lee Sang Hoon, *Yoo Kwan Hee, **Park Kwang Roh
USINGTECH, *CHUNGBUK NATIONAL UNIVERSITY, **ETRI

inteck2@usingtech.co.kr, *khyoo@cbnu.ac.kr, **krpark@etri.re.kr

Abstract. The productivity of nonferrous metals is technically limited. This paper is to approach the process improvement through big data analysis to maximize the efficiency of nonferrous metal manufacturing process and electrical energy. This new approach can respond to future industries by achieving productivity improvement at low cost and comprehensively accepting factors that have not been physically improved

1 Introduction

Through the analysis of big data, one of the key elements of the fourth industry, the company aims to innovate non-ferrous manufacturing processes through logical improvement rather than physical improvement.

Most of the process improvements so far have been made through facilities or direct control S/W, but a more efficient system can be developed by controlling the process by considering indirect elements (temperature, mining conditions, etc.) through big data analysis. If these efforts are successful, AI will be available in the future.

2 Body

2.1 Importance of Big Data Application in Submerged Arc Furnace

Submerged Arc Furnace (SAF) is unable to determine the heating status, ore melting or electrode location, due to its unique heating structure. Thus far, much has been done in a way that depends on the experience of the operator. This resulted in a higher dependency on humans than a dependency on systems, and the development of SAFs was difficult to achieve. In addition, since SAF does not have many sensor signals to measure, it is difficult to make a very simple and complete system.

The author solved these problems through big data analysis and improved the system. Due to the application of big data, indirect data, which lacked physical data and was difficult to identify, were also comprehensively linked to the process.

2.2 Adjustment of power quality and power input

3 The input power was adjusted by linking the voltage, current and power factor data of Phase with the ore phase in the Furnace up to the process stage. This is similar to the conventional method, but more efficient and improved thermal efficiency was achieved by analyzing big data considering the moisture content of the ore, loading time and climate during mining. What makes this achievement even more meaningful is that we can further apply indirect data to increase the number of data in the future and analyze its impact.

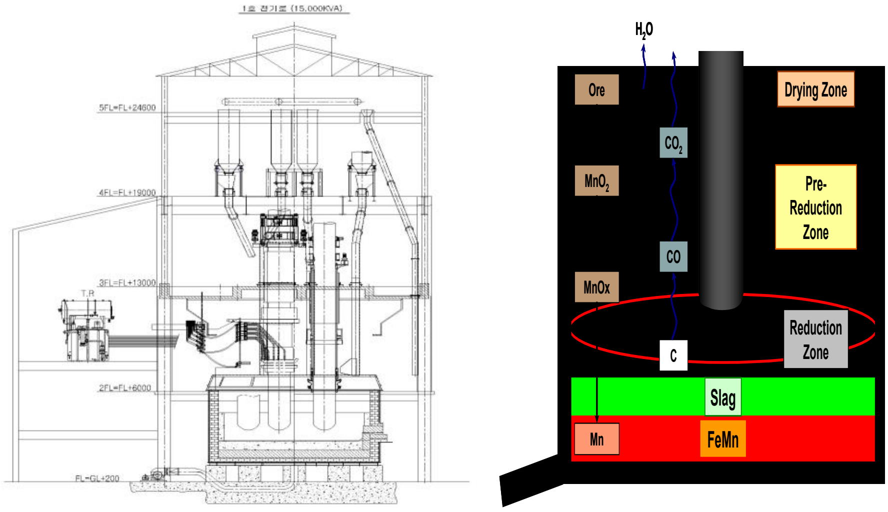


Figure 1

As shown in **Figure 1**, SAF has a very small sensor data despite being a very large installation and electrode rods are buried in the raw material, making it difficult to measure the location of the arc or observe the arc conditions.

2.3 Connecting environmental data around Furnace to big data

The conditions under which nonferrous metals are Melt have many variables, and therefore the energy consumed by the same amount of ore can vary a lot. Until now, the company has mainly used electric energy by observing the conditions of the Furnace, but data from the mining stage of the ore is aggregated and analyzed for big data to ensure proper energy input.

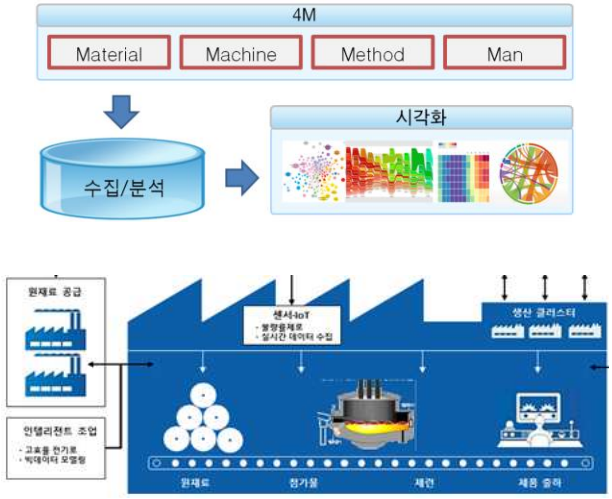


Figure 2

3 Conclusion

Data that has been omitted so far unless it is a direct Furnace factor, such as mining, transportation, storage, time in the Melt phase, meteorological conditions and physical characteristics.

However, considering these factors, it is now possible to develop a more efficient system by developing process processes through big data analysis. In particular, this type of system is significant in breaking away from the system that has been limited to facilities. It is also expected that the development of more logical software is possible and that the portion of software in the overall equipment will increase.

4 Future plans

Develop a new type of system that combines big data with process control system, and then go through a stabilization process to involve more elements. After completing this process, I think it is appropriate to introduce artificial intelligence as the next step.

This paper, which is a step ahead of that, reflects the external and indirect factors that have so far focused on process optimization and cost reduction inside the facility, will further develop the system considering more indirect factors in the future.

A Study on the Necessity of Thermal Analysis in Mechanical Industry Using Thermal Analysis Simulation Program

Sung-Hoon Kim, *Sang-Hoon Lee, **Seong-Bae Seo
PASOL, * USINGTECH, **USINGTECH

fasolsh@daum.net, *inteck2@usingtech.co.kr, **sbseo@usingtech.co.kr

Abstract. In this research paper, using SimsCale thermal analysis simulation program, it is possible to simulate the flow of heat from a physical object and study how reliable and optimal data can be obtained. For example, if you consider a smartphone that warms up when you use a cell phone for a long time, manufacturers need to keep the phone from overheating and make users comfortable, which is one example of having to perform heat transfer simulations to design a safe product.

1 Introduction

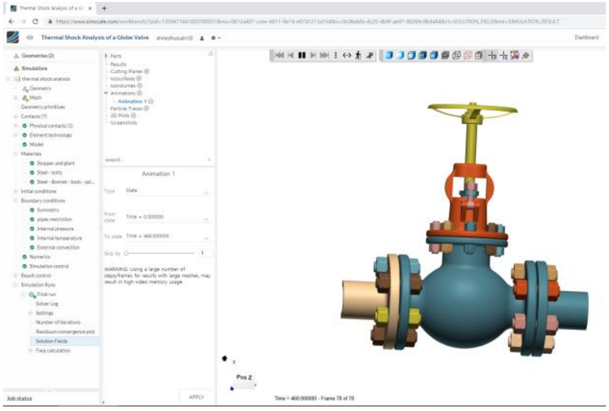
Many products depend on the temperature. Heat may occur and the performance of the product may decrease, and may not continue to operate. For this reason, heat should be designed in the early stages of design to analyze and optimize the product.

Thermal analysis refers to a numerical analysis method that verifies the change and stability of the effects of heat generated by such structures. SimsCale's thermal analysis simulation program provides an analysis of thermomechanical analysis and heat transfer, enabling designers to make optimal designs for thermodynamics and heat transfer.

2 Body

Thermal structural analysis can be divided into thermomechanical analysis and heat transfer analysis types.

Thermomechanical analysis types can calculate the structure and heat behavior of a structure. The results also show the problems that may arise from heat, and the effects of heat on the load condition of the structure.



Animation 1

Type State

From state Time = 0.000000

To state Time = 466.000000

Skip by 1

WARNING: Using a large number of steps/frames for results with large meshes, may result in high video memory usage.

Time = 466.000000 - Frame 70 of 70

Figure 1. Before Globe valve Thermal Impact

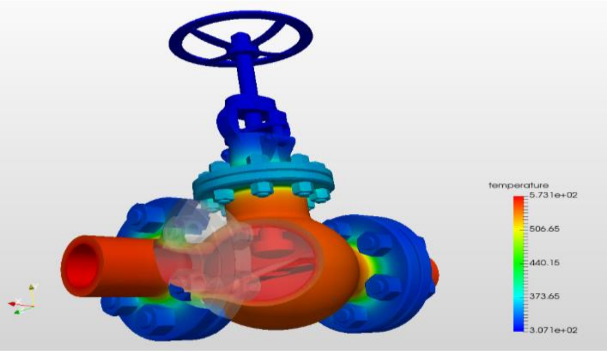


Figure2. After the Globe valve heat shock.

Heat transfer analysis type can simulate heat transfer combined in a fluid through convection and heat transfer in a solid through conduction. Heat transfer generally represents heat flow due to temperature differences and subsequent temperature distribution and variation. The reason for this heat analysis is to verify during the design phase that heat can be safe on the structure when the structure is actually constructed.

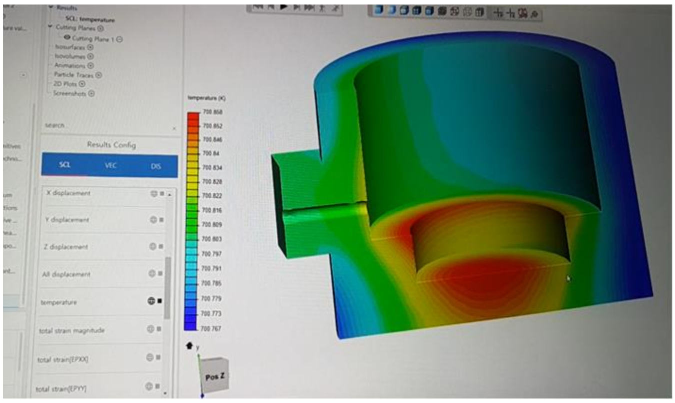


Figure 3. Thermal effects of surfaces according to internal temperature by electric furnace

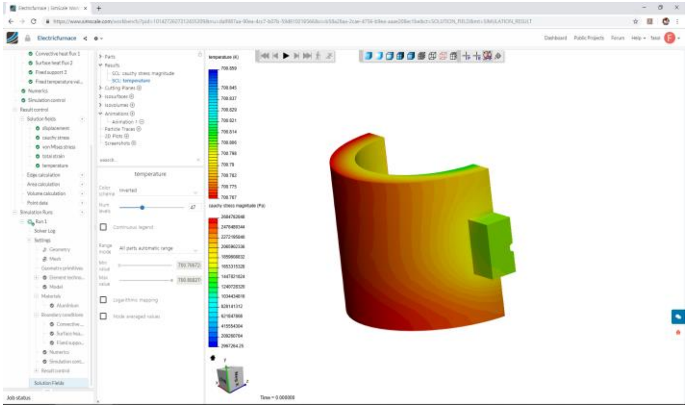


Figure 4. Thermal effects of electric furnace on the interior according to the external temperature

Heat transfer is divided into three types: conduction, convection, and radiation. Conductance means the transfer of heat energy from the hot zone to the cold side, and heat transfer between materials that come into direct contact with each other. Because heat and energy are important variables, it is important to simulate and design the conduction between different materials.

$$Q = KA\left(\frac{dT}{dx}\right)$$

Fourier's law

Convection refers to a heat transfer type that occurs between fluids that flow adjacent to the surface of a solid due to a difference in density when there is a difference in temperature within a fluid. The formula for convection follows Newton's cooling laws.

$$dT/dt = -k(T-S)$$

Newton's cooling laws

Where T is the temperature of the object and k is the constant obtained by the initial condition.

Radiation refers to the phenomenon that the heat energy possessed by an object is released in the form of light-like electromagnetic waves by the thermal motion of the atomic group inside the object, and is one of the methods of heat transfer.

3 Conclusion

Heat analysis requires knowledge of conductivity, convection and radiation, the form of heat transfer, and the flow of heat transfer should be able to prepare designers for heat load effects and failures when setting mechanical components for heat. When structural analysis is performed using thermal analysis simulation software, the temperature distribution can be checked by giving the resultant heat condition to the surface of the structure, or the deformation due to the distributed heat can be interpreted. As such, thermal analysis is a must in the machinery industry and is essential. Designing using heat analysis simulation software will enable designers to predict, analyze, and optimize product heat generation failures early in the design phase to design in a stable structure for proper performance.

4 Reference literature

[1] A Study on the Fire Resistance Behavior of Fiber Mixed High Strength Concrete by Heat Transfer Analysis
[2] Heat Transfer Characteristics in Combustion Nozzle
[3] A Study on the Heat Flow Characteristics of an Outer Rotary Induction Motor using CFD
[4] Thermal Analysis of Welded Coated Continuous Casting Molds
[5] Analysis types based on CFD approach by SimsCale
[6] wikipedia

Enhancement of Geometric One-Class Classifiers

Do Gyun Kim[1] and Jin Young Choi[1]

[1] Department of Industrial Engineering, Ajou University, Suwon, Korea
{rlaehrbs90, choijy}@ajou.ac.kr

Abstract. In this paper, we tackle One-Class Hyper-Rectangle Descriptor ($1-HRD$) which is a state-of-the-art OCC classifier and remedies trade-off between classification accuracy and interpretability of classification results. Specifically, we suggest a novel method for generation of $1-HRD$ that can reflect density and distribution of training data, namely $1-HRD_d$. Furthermore, we design a genetic algorithm for systematical generation of $1-HRD_d$ to tune parameters such as the number of distributions assumed for dataset. Our work is validated by a numerical experiment using UCI machine learning dataset compared to well-known baseline methods. As a result, we can prove superiority of $1-HRD_d$ resulted from the proposed genetic algorithm to other OCC algorithms.

Keywords: One-Class Classification (OCC), One-Class Hyper-Rectangle Descriptor, Genetic Algorithm.

1 Introduction

One-Class Hyper-Rectangle Descriptor ($1-HRD$) is a novel One-Class Classification (OCC) approach that can provide both prominent classification accuracy and interpretability to user, which have been considered as a trade-off in other OCC algorithms[5]. However, existing $1-HRD$ have limitations that they do not consider density or distribution, which might be important for learning patterns of target class. In addition, they require exhaustive and inefficient search procedure for parameter tuning. Based on these motivations, we propose an efficient method for generating $1-HRD$ called $1-HRD_d$ (i) considering density and distribution of data and (ii) resolving inefficient parameter tuning issue by application of Genetic Algorithm (GA).

The rest of this paper is organized as follows. Section 2 reviews for related works on OCC algorithms and Section 3 suggests $1-HRD_d$ using GA. In section 4, a numerical experiment is carried out to validate the performance of the proposed OCC classifier. In the end, we conclude this paper and address further works in Section 5.

2 Related works

There have been many previous works for developing OCC algorithms. Table 1 summarizes their basic concepts and limitations.

Table 1. Information various OCC algorithms.

Methodology	Basic concepts	Pros. and Cons.
Density estimation-based OCC[1,2]	Calculating PDF value of instance	They are easy to implement but threshold, for discriminating class is ambiguous
Decision tree (DT)-based OCC[3,4]	Formulating DT with rules defining class	They extract rules that can be interpreted, but require artificial or unlabeled data
Decision boundary-based OCC[7,9,10]	Finding boundary including instance	They achieve prominent accuracy, but resulted classifiers are too complex
One-Class Hyper-Rectangle (H-RTGL) Descriptor[5,6]	Obtaining H-RTGL surrounding instance	They provide both interpretability and accuracy, but ignore density or distribution and require parameter tuning

3 Suggestion of $1-HRD_d$ using Genetic Algorithm

Chromosome structure and population size. We devise a chromosome as a vector with size q and each element represents the number of Gaussian distributions assumed for each feature, using real-valued encoding scheme. In other words, r-th gene of chromosome represents the number of Gaussian distributions NG_r assumed for feature r. The population size is defined as P, and initial population is generated by randomly. By performing pre-experiment, we set $P=20$.

Generation of classifier and fitness function. At first, we generate intervals corresponding to edges of H-RTGL and obtained feature by feature. Interval generation of $1-HRD_d$ begins with separating projection points in feature r into k_r clusters, where k_r is the value of r-th gene in the considered chromosome. Gaussian distribution $G_r^{q_r}$ corresponding to $q_r(q_r=1,2,\cdots,k_r)$-th cluster of projection points in feature r has mean $\mu_r^{q_r}$ and standard deviation $\sigma_r^{q_r}$, which is used to calculate interval $ITVL_r^{G_r^{q_r}}$ obtained from the Gaussian distribution $G_r^{q_r}$ as follows.

$$ITVL_r^{g_r^{q_r}} = \left[\mu_r^{q_r} - N_\sigma \cdot \sigma_r^{q_r},\ \ \mu_r^{q_r} + N_\sigma \cdot \sigma_r^{q_r}\right], \qquad (1)$$

where N_σ is a parameter to determine the length of interval by controlling the degree of reflection of $\sigma_r^{q_r}$. Then, H-RTGLs are formulated by applying the conjunction of these intervals feature by feature with q-fold Cartesian product operation. Since there are k_r intervals for each feature, maximally $\prod_{r=1}^{q} k_r$ interval conjunctions can be possible. However, considering all interval conjunctions is not efficient and it may generate meaningless intervals including no instance. Resulted interval conjunctions are used as classifiers which have shave of H-RTGLs after adjusting their volume with additional fitting procedure.

Also, as a fitness function, we consider Area Under ROC Curve (AUC) of classifiers obtained by using the information of chromosome. Measuring classifier with AUC can assess performance of the classifier in terms of the ability of excluding outliers as well as including instances of target class.

Crossover Operator. We adopt arithmetical crossover operator that generates offspring chromosomes by arithmetic operation among parental chromosomes. Specifically, we define two gene values y_r^1 and y_r^2 of offspring by using x_r^1 and x_r^2 from parental chromosomes in feature r, which is depicted as follows.

$$y_r^1 = \alpha x_r^1 + (1-\alpha)x_r^2$$
$$y_r^2 = (1-\alpha)x_r^1 + \alpha x_r^2 \quad (2)$$

Mutation Operator. We use a simple integer vector mutation that can be suitable for integer-valued encoding. Specifically, we increase or decrease the value of gene by 1 if it is selected for mutation. This occurs with probability p_m, which is set to $p_m = 0.01$ by pre-experiment.

Population Update and termination condition. After producing new generation with size P by crossover and mutation operations, we select only P chromosomes with high fitness value among $2P$ chromosomes to preserve the size of population. Moreover, as criteria for termination of algorithm, we suggest stopping the algorithm if there is no improvement of solution through the number of generations (exactly 5). Since this policy uses computation resources more efficiently than other conditions [8].

4 A Numerical Experiment

We designed a numerical experiment with datasets provided by UCI machine learning repository, namely Iris, Breast, Biomed, Liver. Since there exist multiple classes in those datasets, we chose one class as target class and other classes as outlier. Moreover, we used 50% of instances belonging to target class as training dataset to learn $1-HRD_d$, whereas the rest of instances in target class and outliers were used as test dataset to verify the classifier. For performance measure, we considered AUC that can assess $1-HRD_d$ in terms of the ability of excluding outliers as well as including instances of target class.

Table 2 displays an experimental result of 20 replications used to measure classification performance of proposed classifier generated by GA compared to other OCC algorithms whose results were taken from references. [6,11]. We indicate the class used as target class in heading of each column.

Table 2. Comparison of AUC obtained from $1-\mathrm{HRD}_d$ and other OCC algorithms

	Iris (Virginica)	**Breast (Malignant)**	**Biomed (Normal)**	**Liver (Healthy)**
	AUC * 100 (standard deviation)			
$1-\mathrm{HRD}_d$ (with GA)	98.2 (1.0)	96.2 (0.8)	90.3 (1.1)	61.1 (1.4)
$1-\mathrm{HRD}_m$	96.1 (1.1)	95.1 (1.2)	89.6 (1.2)	61.6 (1.6)
$1-\mathrm{HRD}_p$	97.6 (0.9)	96.2 (0.7)	85.2 (1.2)	59.4 (2.3)
Naïve Parzen	95.4 (1.1)	96.5 (0.4)	93.1 (0.2)	61.4 (0.7)
Gauss	97.8 (0.6)	82.3 (0.2)	90.0 (0.4)	58.6 (0.5)
SVDD	98.1 (0.8)	70.0 (0.6)	2.2 (0.3)	4.7 (1.4)

We could observe that the classification accuracy of $1-HRD_d$ was better than or similar to that of existing $1-HRD$s in most of datasets, except for Liver dataset. These results support that there exists obvious improvement in $1-HRD_d$, although more datasets should be considered to confirm the superiority of it. Although performance of $1-HRD_d$ was slightly lower than Naïve Parzen classifier, it can provide interpretability of classification results important for post analysis of dataset.

5 Conclusion

In this paper, we proposed an efficient OCC classifier $1-HRD_d$ by using GA. Especially, we devised encoding scheme that could represent the number of Gaussian distributions assumed in each feature r and other operators such as crossover and mutation. As a result, we could achieve desirable classification performance compared to other OCC algorithms as well as existing $1-HRD$s.

Also, we suggest some topics for further research such as more complex and larger dataset, population-based metaheuristics other than GA, and other interval generation methods.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2017R1A2B4009841).

References

1. Agusta, Y., Dowe, D. L.: Unsupervised learning of gamma mixture models using minimum message length. In: Proceedings of the 3rd IASTED Conference on Artificial Intelligence and Applications, pp. 457-462. ACTA Press, Benalmádena (2003)
2. Barnett, V., Lewis, T.: Outliers in statistical data. Wiley, New York (1994)
3. De Comite, F., Denis, F., Gilleron, R., Letouzey, F.: Positive and unlabeled examples help learning. In: International Conference on Algorithmic Learning Theory, pp. 219-230. Springer, Heidelberg. (1999)
4. Denis, F., Gilleron, R., Letouzey, F.: Learning from positive and unlabeled examples. Theoretical Computer Science, 348(1), 70-83 (2005)
5. Jeong, I. K., Choi, J. Y.: Design of One-Class Classifier Using Hyper-Rectangles. Journal of Korean Institute of Industrial Engineers, 41(5), 439-446 (2015)
6. Jeong, I., Kim, D. G., Choi, J. Y., Ko, J.: Geometric one-class classifiers using hyper-rectangles for knowledge extraction. Expert Systems with Applications, 117, 112-124 (2019)
7. Scholkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., Platt, J. C.: Support vector method for novelty detection. In: Advances in neural information processing systems, pp. 582-588. MIT Press, Denver (2000)
8. Sivanandam, S.N., Deepa, S.N.: Introduction to genetic algorithms, Springer, Heidelberg (2008)
9. Tax, D. M., Duin, R. P.: Support vector domain description. Pattern recognition letters, 20(11-13), 1191-1199 (1999)
10. Tax, D. M., Duin, R. P.: Support vector data description. Machine learning, 54(1), 45-66 (2004)
11. Tax, D.M.J., One-class classifier results, http://homepage.tudelft.nl/n9d04/occ/

Impact of Image Pre-processing on Feature Correspondences

Muhammad Umer Kakli[1] , Yongju Cho[1,2] and Jeongil Seo[2]

[1] Korea University of Science and Technology, Daejeon, South Korea
[2] Electronics and Telecommunications Research Institute, Daejeon, South Korea
{umar.isb, yongjucho, seoji}@etri.re.kr

Abstract. Features matching is an essential and challenging step in a wide range of computer vision applications. The task becomes even more difficult if the matching image pairs suffer from distortions such as blurriness or uneven illumination. The performance of feature matching algorithms decreased drastically in such scenarios. Therefore, in this paper, we propose to pre-process the input images such that the maximum number of matching correspondences can be obtained. We first modify the image intensities by applying the Retinex theory based algorithm Low-light IMage Enhancement via Illumination Map Estimation (LIME). Next, scaling is applied to intensity modified images to further improve the number of matched correspondences. The proposed pre-processing scheme is tested with various images of different resolution, brightness, lightening, viewpoint changes, and structural information. The results demonstrate that the number of matched features and epipolar correspondences are significantly increased. On average, the number of matched features and epipolar correspondences are increased by ~144% and ~108%, respectively.

Keywords: Image pre-processing, feature matching, intensity modification

1 Introduction

The detection, description, and matching of image features is extensively used in a number of image processing and computer vision applications, such as image segmentation, object tracking, image stitching and 3D reconstruction. The feature detection and description algorithms intent to represent the image salient information with a small features vector for faster processing. Ideally, these features should be invariant to the illumination, scale and viewpoint changes. In the course of time, various algorithms have been proposed to detect the visually salient features which have shown robustness under above-mentioned distortions. Lowe [1] proposed a Scale Invariant Feature Transform (SIFT) to detect and match the salient image features by exploiting scale-space approach. The detected SIFT features are invariant to the scale, rotation, and partially invariant to illumination and viewpoint changes. Over the time, variants of SIFT have been proposed to handle various other distortions [2]. Regardless of the robustness, SIFT is often not recommended for real-time applications due to its higher computational cost. To this end, Bay et al. [3] proposed Speed-Up Robust Features

(SURF) which improves the speed of detection by utilizing the integral images and box filtering. To further meet the mobile processing requirements binary detectors and descriptors have also been proposed in literature [4].

The goal of feature detection algorithms is to find more number of distinct image features. However, there exist scenarios where even robust feature detection and matching algorithms are unable to provide accurate or enough number of matched features. This may affect the performance of applications where the number of matched features plays a vital role to get acceptable results, such as 3D reconstruction or image stitching. Image pre-processing prior to corresponding matching can be helpful in such cases. It is also evident from the existing literature that pre-processing using image enhancement techniques is capable of improving the performance of feature matching for various applications [5]–[7].

In this paper, we propose a simple yet an effective technique to pre-process the input images to acquire more number of feature correspondences. The proposed scheme comprises of intensity transformation using Low-light IMage Enhancement via Illumination Map Estimation (LIME) algorithm [8] followed by the image scaling operation. LIME along with scaling amplifies the meaningful image information, which helps in identifying more number of correspondences in pair of images. The results demonstrate that the number of matched features are increased by ~144%, while the number of epipolar correspondences are increased by ~108% after application of the proposed pre-processing scheme.

The rest of the paper is organized as follows: In Section 2, the proposed pre-processing scheme is described in detail. In Section 3, we present the experimental details and results. Finally, we provide conclusions in Section 4.

2 Image Pre-processing prior to Correspondence Matching

Histogram Equalization (HE) and Gamma (γ) transformation are the two popular choices to enhance the image contrast due to their computational efficiency. However, for the images with dark content HE results in washed-out artifacts. On the other hand, being effective on dark image contents, the performance of γ transformation on images spanning middle part of the dynamic range is not adequate. Therefore, we used the image enhancement algorithm called LIME [8]. It generates per pixel illumination maps based on Retinex model, which is further refined using structure-aware smoothing model to improve the illumination consistency. Hence, results in better image enhancement as compared to HE and γ transformation.

Let I_1 and I_2 are the two images of a pair to be matched for correspondences. The images after application of LIME are denoted as I_1^{LIME} and I_2^{LIME}, respectively. The objective is to find the optimal scale value at which the maximum correspondences are obtained. Let f be the feature matching function, which returns the optimal scale value when image intensities are modified using LIME, and is defined as:

$$\overset{*}{s} = \underset{a<s<b}{argmax}\, f\left(s(I_1^{LIME}), s(I_2^{LIME})\right) \tag{1}$$

where I_1^{LIME} and I_2^{LIME} are the input images after LIME, $s()$ is the scaling operator, and a and b are the bounds for scaling. Keeping in view the computational overhead, we empirically chosen the values of a and b as 0.5 and 2.5, respectively.

Table 1. Average number of correspondences before and after pre-processing

Sequence	Image Resolution	Average Number of Matched Features		Average Number of Epipolar Correspondences	
		Original	Proposed	Original	Proposed
Bark	765×512	59.4	132.4	34.6	68.8
Bikes	1000×700	359.2	566.6	218.2	336.8
Boat	850×680	240.2	518.2	99.8	193.2
Bricks	1000×700	422.8	1433.0	293.3	768.8
Cars	921×614	170.8	1035.8	129.5	554.3
Graffitti	800×640	210.0	361.0	89.0	119.0
Trees	1000×700	208.4	366.8	108.8	176.6
UBC	800×640	884.2	1767.4	738.8	1350.4
Average	**-**	**319.4**	**772.7**	**214.0**	**445.9**

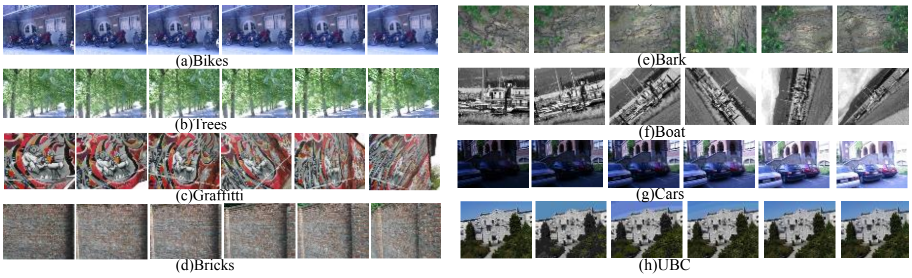

(a)Bikes
(e)Bark
(b)Trees
(f)Boat
(c)Graffitti
(g)Cars
(d)Bricks
(h)UBC

Fig. 1. Mikolajczyk dataset

3 Experimental Details and Results

SURF is widely used due to its robustness and ability to compute the distinctive image features in a computationally efficient way. Therefore, we chose SURF as a feature matching algorithm in our experiments. In addition to feature matching, we also observed the effect of image pre-processing on epipolar correspondences. All experiments are conducted under the same environment using default parameters. We used MATLAB on Intel Core i7 PC with 32GB of RAM for the experiments.

In order to validate the proposed scheme, we used the well-known Mikolajczyk dataset [10] for our experiments, as it offers images with various types of distortions such as blur, viewpoint changes, etc. The dataset consists of eight sequences with different scene information as shown in Fig. 1. The images of the first column of Fig. 1 are treated as reference, and are matched with the rest of the images in the corresponding sequence. The number of matched features and epipolar correspondences without any pre-processing are referred to 'Original'. For every pair, we then heuristically find the maximum number of matched features and Eipoplar correspondences by applying different scaling values after LIME and termed it as 'Proposed'. In Table 1, we present the number of matched features and epipolar correspondences for the aforementioned experimental cases. For each image sequence, we provide the average number of matched features and epipolar correspondences from all consecutive pairs. It can be clearly seen that, on average, by incorporating the proposed pre-processing scheme, the number of matched features and epipolar correspondences are increased by ~144% and ~108%, respectively.

Epipolar correspondences play pivotal role in 3D Reconstruction. The more number of epipolar correspondences can better estimate the system parameters which may lead to generate more number of points in 3D sparse reconstruction process. Our results indicate that the proposed method can be effective in increasing the number of correspondences to improve the performance of 3D sparse reconstruction problems.

3 Conclusion and Future Work

In this paper, we closely analyzed the effect of pre-processing on input images prior to find the feature correspondences. We first modify the intensities of input images by applying LIME algorithm and we then determine the optimal scale to obtained maximum number matched features and epipolar correspondences. Experimental results on various types of images proved that the proposed pre-processing scheme can significantly improve not only the number of matched features but also the epipolar correspondences. On average, the numbers of matched features and epipolar correspondences are increased by ~1.44 and ~1.08 times, respectively.

Future work includes modeling of heuristic process to determine the optimal scale and incorporation of the proposed scheme in 3D reconstruction pipeline to obtained more complete 3D sparse reconstruction model.

Acknowledgments. This work was supported by Korea Government (MSIT) 19ZR1120 (Development of 3D Spatial Media Core Technology).

References

1. Lowe, D. G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. 60(02), 91–110, (2004)
2. Wu, J., Cui. Z., Sheng, V. S., Zhao, P., Su, D., Gong, S.: A Comparative Study of SIFT and its Variants. Measurement Science Review. 13(3), 122–131, (2013)
3. Bay, H., Ess, A., Tuytelaars, T., Goolm, L. V.: SURF: Speeded Up Robust Features. Computer Vision and Image Understanding. 110(3), pp. 346–359, (2008)
4. Heinly, J., Dunn, E., Frahm, J. M.: Comparative Evaluation of Binary Features. European Conference on Computer Vision. (2012)
5. Tu, L., Dong, C.: Histogram equalization and image feature matching. International Congress on Image and Signal Processing. (2013)
6. Su, H., Wang, J., Li, Y., Hong, X., Li, P.: An Algorithm for Stitching Images with Different Contrast and Elimination of Ghost. International Symposium on Computational Intelligence and Design (ISCID). (2014)
7. Wang, T., Lu, G., Xia, Y.: An efficient pre-processing method for feature based image stitching. International Conf. on Wireless Communication and Signal Processing. (2016)
8. Guo, X., Li, Y., Ling, H.: LIME: Low-light image enhancement via illumination map estimation. IEEE Transactions on Image Processing. 26(2), 982–993, (2017)
9. Mikolajczyk Dataset, http://www.robots.ox.ac.uk/~vgg/research/affine/

Data mining from ship AIS base station database

Sang Lok Yoo[1], Keon Myung Lee[1]

[1] Depratment of Computer Science, Chungbuk National University, Cheongju, Korea
{sanglokyoo, kmlee}@cbnu.ac.kr

Abstract. Automatic Identification System (AIS) is a valuable navigational aid of the vessels. The AIS information has been used for various purposes. However, not all vessels accurately report all AIS information. Lost messages influence prediction algorithms. In this study, not only the AIS reception rate for class type and vessel type is analyzed, but also shadow zones of the AIS are identified. This study proposes the methods of identification of shadow zones for AIS.

Keywords: Shadow zone, Identification, AIS, Reception rate

1 Introduction

Automatic Identification System (AIS) is a valuable navigational aid, one of several on the bridge of a vessel. AIS provides for transmitting messages between AIS equipment, which can be established on vessels and base stations. The AIS can detect vessels within Very High Frequency (VHF) range around bends and behind islands [1]. The AIS information has been used for various purposes. AIS can be used to identify potential collisions by plotting the course and speed of vessels.

There are two different types of AIS transceivers available. Class A AIS is mandatory equipment, and Class B AIS is for smaller vessels. AIS data includes static, dynamic, and voyage related information. Static information such as vessel name, Maritime Mobile Service Identity (MMSI), call sign, length, breadth, and vessel type is set into the device at commissioning. Dynamic information such as timestamp-date and time, longitude, latitude, course, heading, and speed is synchronized from interfaces with the ship's Global Positioning System (GPS), gyrocompass, and other devices. Voyage-related information such as destination, draft, and Estimated Time of Arrival (ETA) is entered manually by the navigators [2].

However, not all vessels accurately report all AIS information. Lost messages influence prediction algorithms. Although it is important to interpolate the AIS data, it is also important to identify which areas are ultimately shadowed. Interpolations were also affected by the shadow zones [3].

There have been many studies on the identification of radar shadow areas [4 - 8]. However, there are not many studies has yet been conducted to identify AIS shadow zones. Although Shelmerdine [9] shows the overall distribution of the AIS coverage with shadow zones, the methods didn't focus on identification shadow zones of AIS. The output displayed the density grid based on vessel tracks. If the vessels can't navigate due to the aquaculture farms or low depth of the sea in a particular area, no AIS data naturally is collected. There is a possibility that it is judged as shadow zones because there are few ship traffic. Therefore, a method for identifying AIS shadow zones is required even in areas with high vessel traffic volume.

In this study, not only the AIS reception rate for class type and vessel type is analyzed, but also shadow zones of the AIS are identified. In particular, this study proposes the methods of identification of shadow zones for AIS.

2 Materials and methods

2.1 AIS datasets

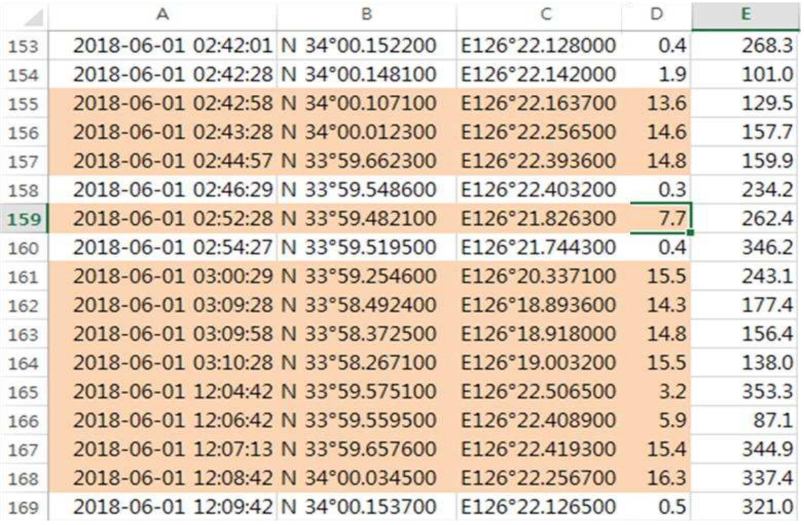
The AIS data was obtained through the AIS national integrated system of the Ministry of Oceans and Fisheries of South Korea (MOF). Between 2001 and 2008, 42 AIS base stations were constructed on the coasts of South Korea. Each AIS base station will have a range of 40 nm. There are two AIS base stations, Heugildo island and Chengsando island, in the study area. The AIS base stations used by the Wando VTS are located in sites that ensure coverage of the VTS monitoring area. The exact location of AIS base station is 34˚12.91́N, 126˚53.17́E at Chengsando island, and 34˚ 16.54́N, 126˚32.43́E at Heugildo island . The AIS data used in this analysis was recorded from June 1 to June 9, 2018. The dataset includes Class A AIS as well as Class B AIS data messages. AIS data were collected from a total of 328 vessels in the Wando VTS area. Class A type AIS data were collected from 236 vessels including 54 fishing vessels, 159 cargo vessels, and 23 passenger vessels. Class B type AIS data were collected from 92 vessels including 63 fishing vessels and 29 leisure fishing boats. The dataset consisted of a colossal amount of dynamic and static information according to the MMSI. The MMSI and vessel type among the static data is used for analyzing, and the dynamic data is used for the identification of AIS shadow zones. The database comprises about 1 million datasets for Class A dynamic messages and 140 thousand datasets for Class B dynamic messages. The percentage of Class B messages in relation to Class A is below 13.4 percent.

2.2 Approach of the data analysis

In the first step, the preprocessing of datasets. The seven AIS variables (MMSI, vessel type, date, time, longitude, latitude, speed) were chosen for the analysis. The AIS variables including such as the longitude, latitude, and speed are rare with unknown and error values, as compared to vessel heading (BANYŚ et al., 2012). AIS message errors regarding the vessel type (Harati-Mokhtari et al., 2007) were corrected using data published by the International Telecommunication Union (ITU, 2014; Yoo, 2018) and the port management information system (Port-MIS) data of the MOF

(Park et al., 2005). Then, AIS messages sent from positions outside the study area were discarded.

In the second step, criteria were set for AIS data loss positions. It is necessary to know the properties and transmission periods of Class A type and Class B type. AIS-equipped vessels periodically broadcast position data. The class A type broadcast very frequently the dynamic data at 2-10 seconds intervals depending on the vessels speed while underway. Whereas class B type has longer reporting intervals than class A. Vessels going less than 2 knots transmit dynamic data every 3 minutes while vessels are navigating more than 2 knots updates dynamic data every 30 seconds at Class B type.



	A	B	C	D	E
153	2018-06-01 02:42:01	N 34°00.152200	E126°22.128000	0.4	268.3
154	2018-06-01 02:42:28	N 34°00.148100	E126°22.142000	1.9	101.0
155	2018-06-01 02:42:58	N 34°00.107100	E126°22.163700	13.6	129.5
156	2018-06-01 02:43:28	N 34°00.012300	E126°22.256500	14.6	157.7
157	2018-06-01 02:44:57	N 33°59.662300	E126°22.393600	14.8	159.9
158	2018-06-01 02:46:29	N 33°59.548600	E126°22.403200	0.3	234.2
159	2018-06-01 02:52:28	N 33°59.482100	E126°21.826300	7.7	262.4
160	2018-06-01 02:54:27	N 33°59.519500	E126°21.744300	0.4	346.2
161	2018-06-01 03:00:29	N 33°59.254600	E126°20.337100	15.5	243.1
162	2018-06-01 03:09:28	N 33°58.492400	E126°18.893600	14.3	177.4
163	2018-06-01 03:09:58	N 33°58.372500	E126°18.918000	14.8	156.4
164	2018-06-01 03:10:28	N 33°58.267100	E126°19.003200	15.5	138.0
165	2018-06-01 12:04:42	N 33°59.575100	E126°22.506500	3.2	353.3
166	2018-06-01 12:06:42	N 33°59.559500	E126°22.408900	5.9	87.1
167	2018-06-01 12:07:13	N 33°59.657600	E126°22.419300	15.4	344.9
168	2018-06-01 12:08:42	N 34°00.034500	E126°22.256700	16.3	337.4
169	2018-06-01 12:09:42	N 34°00.153700	E126°22.126500	0.5	321.0

Fig. 1. Sample of ship AIS database. The ‘A’ column is data and time. The ‘A’ column is date and time. The 'B' is longitude and the 'C' is latitude. The 'B' is longitude and the 'C' is latitude. The 'D' is speed and the 'E' is course.

The reporting interval with the date and time variable in each dynamic data rows was calculated in seconds. The reporting interval is measured between two sequenced position reports of a vessel rounded in seconds. At least two consecutive dynamic data were extracted with a speed of 2.0 knots or more. It is because that when Class B type vessels are navigating with more than 2.0 knots, they should update dynamic data every 30 seconds. If there is a reporting interval of more than 10 minutes (=600 seconds) among the extracted data, it should be separated into other groups. Fishing vessels and leisure fishing boats equipped with Class B type are reluctant to position exposure due to the nature of their operation. Therefore the AIS power is frequently forcibly cut off, and dynamic data loss is assumed to be frequent. In this study, only the data with a reporting interval of fewer than 10 minutes was used as the analytical data, assuming 10 minutes as the threshold value. Data exceeding the threshold is a message loss caused by the forced power down because it can affect the shadow zones we are trying to identify in this study. Then the first row in each group is deleted because no reporting interval can be calculated. The data from each group is agglomerated to form the final data output.

3 Results and Discussion

VTS watchkeepers must not assume that the collision avoidance solutions proposed by the AIS equipment are necessarily accurate or correct.

Acknowledgements

This work was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea, Republic of Korea (Grant no.: NRF-2017M3C4A7069432).

References

1. Last, P., Hering-Bertram, M., & Linsen, L. (2015). How automatic identification system (AIS) antenna setup affects AIS signal quality. Ocean Engineering, 100, 83-89.
2. International Maritime Organisation (IMO), 2002. Guidelines for the Onboard Operational Use of Shipborne Automatic Identification Systems (AIS), Assembly 22nd Session, Resolution 917.
3. Shelmerdine, R. L. (2015). Teasing out the detail: How our understanding of marine AIS data can better inform industries, developments, and planning. Marine Policy, 54, 17-25.
4. Bennett, A. J., & Blacknell, D. (2003). The extraction of building dimensions from high resolution SAR imagery. In Radar Conference, 2003. Proceedings of the International (pp. 182-187). IEEE.
5. Eineder, M., & Suchandt, S. (2003). Recovering radar shadow to improve interferometric phase unwrapping and DEM reconstruction. IEEE Transactions on Geoscience and Remote Sensing, 41(12), 2959-2962.
6. Cui, J., Gudnason, J., & Brookes, M. (2005). Radar shadow and superresolution features for automatic recognition of MSTAR targets. In Radar Conference, 2005 IEEE International (pp. 534-539). IEEE.
7. Prasath, V. S., & Haddad, O. (2014). Radar shadow detection in synthetic aperture radar images using digital elevation model and projections. Journal of Applied Remote Sensing, 8(1), 1-10.
8. Lu, Z., Wang, Y., Yuan, Y., Wei, Y., & Huang, Y. (2017). Research on retrieving wave height from radar shadow images based on wind angle feature. In Mechatronics and Automation (ICMA), 2017 IEEE International Conference on (pp. 1092-1097). IEEE.
9. Shelmerdine, R. L. (2015). Teasing out the detail: How our understanding of marine AIS data can better inform industries, developments, and planning. Marine Policy, 54, 17-25.

A Relation Analysis Using Sound and Posture for Preventing PRMDs

So-Hyun Park[1], Sun-Young Ihm[1], Mi-Yeon Kim[2], Young-Ho Park[1,*]

[1] Dept. of IT Engineering
Sookmyung Women's University,
[2] Dept. of Airline Services & Secretarial Studies
Seoyeoung University,
{shpark, sunnyihm, yhpark}sm.ac.kr
myk@seoyeong.ac.kr

Abstract. It is essential to prevent playing-related musculoskeletal disorders (*PRMDs*) of piano players as it is directly linked to the players' performance. Over the years, there have been several studies that focus on recognizing correct postures and prevent injuries. Unlike previous studies, this paper proposes a multi-modal based injury prevention study that deals with audio and visual data. Specifically, in this paper, we study the relationship of sound to the posture and how the deep learning model can be used to prevent injuries of piano players using sound. The initial experiment results demonstrate the accuracy of the proposed method when recognizing incorrect postures.

Keywords: multi modal learning, Playing-realted musculoskeletal disorders, sound analysis, timbre learning, posture analysis, deep convolutional neural network

1 Introduction

Playing-related musculoskeletal disorders (PRMDs) are injuries that can significantly affect the piano players' performance and future professional life. There have been several studies to prevent such injuries of piano players. Here, recognition and analysis of the posture are essential since it is the first step to prevent injuries in advance. Thus, most of the related studies are being conducted to define and recognize various postures to prevent injuries for playing such as tennis or beach volleyball and so on. There are also studies being conducted to prevent injuries can be occurred in handling musical instruments as well as in physical playing or education [1-8].

Although there is evidence that posture and sound are related when playing music, there is a lack of research that reveal the relationship between players' posture and its'

[1] Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

sound through computational experiments. Unlike previous studies that deal with only visual data to prevent players' injuries, this paper proposes a novel injury prevention study based on multi-modal methods. Specifically, we investigate the relationship between sounds and visual features in playing music of the video data. In addition, we apply a well-known deep learning model, namely deep convolutional neural network (*DCNN*) that can be used to train data and reveal the injury-related postures.

The rest of this paper is organized as follows. Section 2 describes the existing work related to this paper. Section 3 explains relation analysis model between injury-related posture and sound. Section 4 conduct an experiment and analyze the results. Section 5 summarizes and concludes the paper.

2. Related Works

Recently, multi-modal research has become active as emerging various and heterogeneous big data. Among multi-modal studies related to visual and auditory perception, there is shown in research of Zhao et al. The research have been studying the location of sound by using visual and auditory information in the performance video [1]. Gao et al. also researches to separate the sound of each instrument in a video that plays two or more instruments [2]. Ephrat et al. proposed a study to extract each human voice from a video in speech of two people [3].

In previous studies related to injury preventions, the fields which mainly carried out are art, music and physical education and so on. Sohyun et al. have conducted research to detect injury-related postures to prevent injury when playing the piano [4]. Kautz et al. proposed an automatic monitoring system for the prevention of injuries in beach volleyball [5]. Similarly, Mora et al. conducted a study to classify tennis movements by recognizing tennis actions [6].

The timbre of the sound is a complicated factor that can make two sounds feel different when pitch and intensity are the same [7]. There are many studies to classify timbre of the sound. Aucouturier et al. proposed a timbre model which search and analyze music signals [8]. The proposed model can be used to calculate the similarity between songs or to extract repeating patterns in meaningful or efficient within a song. Pons et al. proposed a study on the effective modeling of timbre from log-mel magnitude spectrogram using a deep convolutional neural network [9] (fig 1.).

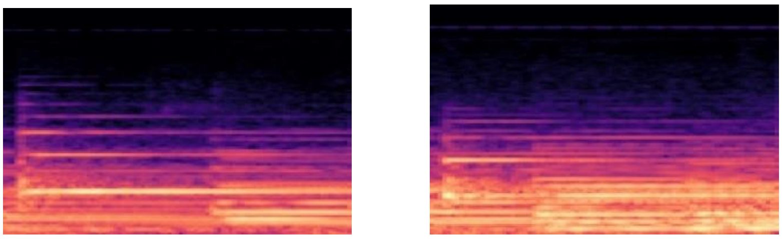


Fig. 1. Input data (mel spectrogram)

3. Relation Analysis Model

In this paper, we use a DCNN to analyze the relationship between the injury position and the sound timbre. If the deep learning model classifies the two sounds well, it is assumed that there is a relationship between sound and posture. Our analysis model uses the architecture like in the below reference, [10]. Kernel size was (3, 3) and pool size was (2, 2). The filter number of convolutional layer was 32. Input shape was (*channel*, *n_mels*, *time*). *n_mels* means number of mel bins. time means audio duration which is split into three seconds. The first layer of *DCNN* was a convolutional layer with activation function *Relu* (*Conv Ⅰ*). The second layer of Architecture was a 4 iteration of the composition of convolutional layer(*Conv Ⅱ~ Ⅴ*) and max-pooling. In this process, this model use Elu activation function and dropout. Percentage of dropout was 0.25. DCNN model use the two fully connected layer with dropout. Percentage of dropout was 0.5.

4. Experiment

In this paper, unlike the previous research, we judge the injury attitude based on sound. If the deep learning model classifies the two sounds well, it is assumed that there is a relationship between sound and posture. In this study, we collect audio data performed by one professional pianist. The collected music piece is C, B, and F major of *Charles-Louis Hanon* (fig. 2). Pianist played the same music piece by two posture (Correct posture and incorrect posture). The correct posture is straight neck with straight waist. The other hand, incorrect posture is bended neck with straight waist (fig. 3). When player postured correctly, audio is stored in correct posture class. Otherwise, when player postured incorrectly, audio is stored in incorrect posture class.

The computer specifications for the development were the Intel® Xeon® CPU E5-2620 v3 @ 2.40GHz and the graphics cards used Titan Xp and GeForce GTX TITAN.

The experimental results are analyzed with three perspectives: Test_loss, Test_acc, AUC. Test_acc means test accuracy. Range of Test_acc is 0~1. Area under curve(AUC) is values, which are often used to score binary classification models. Range of AUC is 0~1. In AUC, close to 1 means higher accuracy.

Experiments were conducted to compare the classification accuracy when data number is changed to 30, 60, 90, and 120. Audio is split into three seconds. We used a mel-spectrogram as an input which express the sound timbre [9]. In the experiment in below, the training, validation and test are divided in 8:1:1 respectively. Early stopping option was given as monitor val_acc. In the experiment, averages of test_loss, test_acc, and AUC were 1.0225, 0.7083, and 0.6528 each. This result show sound is varied with posture and this model can distinguish two types of sound timbre well.

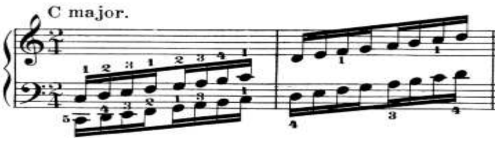


Fig. 2. Example of C major scale

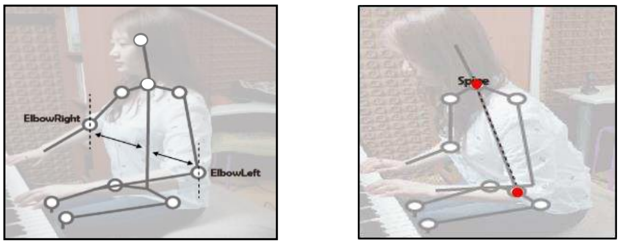


Fig. 3. Correct posture and incorrect posture

5. Conclusion

Multi-modal learning was conducted to prevent injuries can be occurred when playing the piano. Not only conventional piano performance, but also other areas related to the injury prevention research have been focused on only visual data, but in our research, there is a difference in that sound information are additionally utilized. There is a meaningful connection between posture and sound when playing the piano, but this is not steel revealed in the previous experiment. Therefore, in this study, we have tried to collect audio data corresponding to correct posture and incorrect p posture (turtle neck) postures and classify them using DCNN model. As a result, the classification accuracy of 0.8% was obtained. This is because the deep learning model learns well the feature points of the audio depending on the posture. Also, DCNN can classify the class so that there is a relationship between the injury posture and the timbre of the sound in playing piano. However, in the current experiment, it is limited to one song, so it is necessary to perform additional experiments on several songs as further studies expansively, and it is expected that it can be used to improve the accuracy of classifying the injury posture.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2018R1D1A1B07046550). In addition, this work was supported by Institute for Information communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2016-0-00406, SIAT CCTV Cloud Platform).

References

[1] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott and A. Torralba, "The sound of pixels," In Proceedings of the European Conference on Computer Vision (ECCV), pp. 570–586, 2018.

[2] R. Gao, R. Feris and K. Grauman, "Learning to separate object sounds by watching unlabeled video," In Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–53, 2018.

[3] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W.T. Freeman and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," ACM Trans. Graph, vol. 37(4), pp. 1–11, 2018.

[4] S.H. Park, S.Y. Ihm, A. Nasridinov and Y.H. Park, "A Feasibility Test on Preventing PRMDs based on Deep Learning," Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[5] T. Kautz, B.H. Groh, J. Hannink, U. Jensen, H. Strubberg and B.M. Eskofier, "Activity recognition in beach volleyball using a Deep Convolutional Neural Network," Data Mining and Knowledge Discovery, vol. 31(6), pp.1678–1705, 2017.

[6] S.V. Mora and W.J. Knottenbelt, " Deep learning for domain-specific action recognition in tennis," In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 170–178, IEEE, 2017.

[7] T. Letowski, "Sound quality assessment:concepts and criteria," In Audio Engineering Society Convention 87, 1989.

[8] J.J. Aucouturier, F. Pachet and M. Sandler, ""The way it sounds": timbre models for analysis and retrieval of music signals," IEEE Transactions on Multimedia, vol. 7(6), pp.1028–1035, 2005.

[9] J. Pons, O. Slizovskaia, R. Gong, E. Gómez and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," In 2017 25th European Signal Processing Conference (EUSIPCO), pp. 2744–2748, 2017

[10] Scott H. Hawley, "Panotti: A Convolutional Neural Network classifier for multichannel audio waveforms (Version v1.0.0)," Zenodo. http://doi.org/10.5281/zenodo.1275605, 2017.

An Architecture of Reliable and Inexpensive Disk Storage for Data Archiving

[1]Seung Hee LEE, [*]Duseok Jin, Jeong-Heon Kim, Heejune Han, [2]Seo-Young Noh
[1]Korea of Science and Technology Infromation(KISTI), 245, Daehak-ro, Yuseong-gu,
Daejeon, Republic of Korea
{[1]benecia, [*]dsjin, jh.kim, hjhan}@kisti.re.kr
[2]Chungbuk National University, Chungdae-ro 1, Seowon-Gu,
Cheongju, Chungbuk 28644, Republic of Korea
[2]rsyoung@cbnu.ac.kr

Abstract. As the amount of information that we utilize in the future increases dramatically, the size of data center storage devices is expected to increase more and more. On the other hand, in most data centers, the available budget for storage equipment deployment is limited, therefore it is important to find a way to have more storage space with reasonable cost. In this paper, we propose a disk-based storage architecture that secures the maximum storage space and archives data security at the same time.

Keywords: DAS, JBOD, Storage Architecture, Storage Solution

1 Introduction

As the amount of data generated in various layers such as individuals and companies worldwide is increasing year by year, storage devices such as HDD(hard disk drive), SSD(solid-state drive), and tape drive are continuously developled. Nevertheless, as the amount of data generated in the IoT and Big Data era increases beyond imagination, it will be important to secure the maximum storage space within a limited budget

KISTI GSDC(Global Science experimental Data hub Center) provides data sharing services and data analysis environments for domestic and international scientific research professionals conducting data-intensive research such as high-energy physics, nuclear physics, genomics, and electron microscopy. GSDC needed an alternative solution to archival data preservation in the long run as the tape system currently in operation was market-proprietary by a specific manufacturer. The disk is not only the most widely deployed and globally operational storage device, but it can also be a reasonable alternative to long-term data retention, as the prices of disk storage are expected to decline over time. Therefore, we have been studying ways to secure maximum storage space for a cost.

The main purpose of this paper is to present a disk-based archival data storage architecture for KISTI GSDC to support researchers who are studying high energy physics using CERN data. However, the proposed architecture can be used to implement similar large-scale data storage systems as well as service purposes to support high-energy physics.

2 Proposed Architecture

In this type of archival data, the frequency of use of data is low, but the amount of data to be stored is large. For this reason, the size of the storage space is considered more than the performance when building the storage device. Therefore, we propose a method to secure the storage capacity while maintaining the stability of data based on the disk-based storage device that is currently in operation with the largest volume of data.

In order to obtain the maximum capacity for price, the hardware architecture is a DAS(Direct-Attached Storage) structure. DAS is simpler to setup and configure than NAS or SAN. It also has a cheaper price point in terms of raw storage. DAS connects external storage directly to a card connected to the server's internal bus, without the need for an Ethernet or FC(Fiber Channel) switch, such as NAS(Network-Attached Storage) or SAN(Storage Area Network). As a result, the server does not need to browse the network to read and write data, so it can provide users with better performance than network storage. In addition, DAS can easily expand storage capacity by purchasing an external storage enclosure and connecting it to the server, which simplifies the structure and lowers maintenance costs, and costs less than SAN or NAS when deployed.

JBOD(Just a Bunch of Disk / Drives) is a drive enclosure that can hold multiple disk drives. With the JBOD architecture, an administrator can create and organize groups of hard drives of different sizes into a single logical volume or a group of individual hard drives. Drives in the enclosure are connected through a common backplane and can be connected directly to the SAS(Serial-Attached SCSI) HBA(Host Bus Adapter) card on the management server via a SAS cable, allowing for fast and easy expansion of array capacity. JBOD has excellent read speed and write performance, and can be easily constructed. It is also cheaper than other storage systems.

Drives in the enclosure are connected via a common backplane and can be connected directly to the management server via a SAS HBA card, allowing for rapid and easy expansion of array capacity.

On the other hand, JBOD is a simple disk array that does not provide fault tolerance when there is no hardware RAID(Redundant Array of Independent Disks or Redundant Array of Inexpensive Disks) controller, so there is a risk of data loss.

We implemented a software-based RAID method to prevent data loss without an array controller. RAID arrays store data on multiple disks in a way that duplicates data or stripes data across multiple disks in a way that achieves better performance than one disk can provide.

Typically, a software RAID array appears as a single disk in the operating system, and Array management functions are implemented by software running in the host environment, such as an operating system. Software RAID requires additional transfers over the I/O bus. Although software RAID is not as diverse as hardware

RAID, it can reduce reliability and cost with multi-pathing capabilities that support multiple hosts.

When software RAID is set to JBOD, software RAID constructs RAID using block device recognized by OS after OS of server for enclosure management is booted.

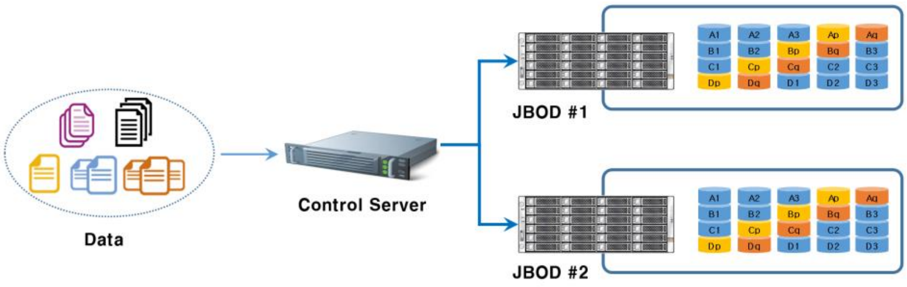


Fig. 1. Proposed Architecture

When software RAID is set to JBOD, software RAID constructs RAID using block device recognized by OS after OS of server for enclosure management is booted.

Table 1. Comparison with existing storage architecture

Categories	SAN	NAS	Proposed Architecture
Price	●●	●●●	●
Storage Capacity	●●	●	●●●
I/O Protocol	SCSI	NFS, CIFS	SCSI
Switch	○ (SAN Switch)	○ (Network Switch)	X
Access Cable	Fibre Channel	Ethernet	HD Mini SAS
Read/Write Speed	●●	●	●●●
Software RAID	X	X	○

Currently, GSDC has two types of disk-based storage architectures, SAN and NAS. We compared the proposed architecture with these two storage architectures. The proposed architecture is cheaper because it does not have expensive hardware controller, so more storage capacity can be achieved with the same cost. It also does not require a SAN switch or a network switch because it is directly cabled to the storage. On the other hand, since the proposed architecture does not have a hardware controller that provides RAID function in a storage device, a software RAID is required to guarantee data security.

3 Conclusion

In this paper, we have presented a method for constructing disk-based storage focusing on KISTI GSDC's data archiving solution.

In the future, the amount of information that we will use will increase significantly, and data centers and big data environments based on them will increase. As a result, the importance of configuring storage solutions to provide more storage space is increasing. In the meantime, data archiving has been mainly performed using tape drives, but the possibility of being dependent on the specific monopolization of certain manufacturers is increasing. As disk storage devices continue to decline in prices, alternatives are needed.

The proposed software RAID based large capacity JBOD architecture can provide data stability and large storage space in this environment. Such a structure would be suitable for data archiving applications that require more data space than performance aspects.

Acknowledgments

This study was supported by the National Research Foundation of Korea (NRF) through contract N-19-NM-CR01-S01 and the Program of Construction and Operation for Large-scale Science Data Center (K-19-L02-C03-S01).

References

1. Andreas-Joachim Peters, Elvin Alin Sindrilaru, Philipp Zigann: Evaluation of software based redundancy algorithms for the EOS storage system at CERN J. Phys. Conf. Ser. 396 042046(2012)
2. R.Castro, L.Abadie, Y.Makushok, M.Ruiz, D.Sanz, J.Vega, J.Faig, G.Román-Pérez, S.Simrock, P.Makijarvi: Data archiving system implementation in ITER's CODAC Core System. Fusion Engineering and Design, vol. 96–97, pp. 751-755, ELSEVIER(October 2015)
3. D. A. Patterson, G. Gibson, and R. H. Katz: A case for redundant arrays of inexpensive disks(RAID). ACM, vol. 17, no. 3. vii, 4, 41, 42(1988)
4. C. Cecchinel, M. Jimenez, S. Mosser, and M. Riveill: An architecture to support the collection of big data in the Internet of Things, pp. 442–449. 1, 2014 IEEE World Congress
5. David Vellante: Scaling Storage for the Cloud: Is Traditional RAID Running out of Gas, http://wikibon.org/wiki/v/Scaling_Storage_for_the_Cloud:_Is_Traditional_RAID_Running_out_of_Gas

Multiple Fracture Classification Using Deep Learning

Sang Hyun Lee[1] , Chan Sik Han[1], Seung Myung Choi[2], Keon Myung Lee[1]

[1] Department of Computer Science, Chungbuk National University, South Korea
[2] Konkuk University Chungju Hospital Orthopedics, South Korea

sanghyunlee@chungbuk.ac.kr, chatter0502@gmail.com
davidchoi1530@gmail.com, kmlee@cbnu.ac.kr

Abstract. The orthopedists use Computed Tomography(CT) to identify characteristics of fracture and determine treatment method. The orthopedists must have a high level of expertise to classify multiple fractures. In this paper, we casts the fracture category classification process as a multi-label classification problem, and proposes a method using deep learning. The proposed model extracts the features of fractures with inception module and classify fracture into each categories based on the computed score. As a result, the proposed method showed 73.3% precision and 86.9% recall in the experiments.

Keywords: Deep Learning, CNN, Multi-Label Classification, Fracture

1 Introduction

The orthopedists use CT images of the fractured patient to identify the anatomical location, shape, and angle of the fracture, and they classify the categories of fracture. This process is essential for the orthopedist determines treatment method by characteristics of fracture. However, statistics of the Health and Welfare of Korea shows the ratio of 'doctors : patients' was '1000 : 2.3' in 2017.[1] In this situation, the orthopedists must be repeat this process to every patients. And fractures is often multiple fracture which have various categories at the same time. For this reason, this process required for orthopedists to have stressful examination and high level of expertise. In this paper, we propose a deep learning based method to resolve these problem. The propose method assist the orthopedists identify fractures and further automate that process. We use CNN to classify multiple fractures based on CT images of lower body fractured patients.

2 Related Work

The experiment conducted by this paper is a multi-label classification problem with one or more labels for each CT image. We use precision and recall as a metric for multi-label classification.[2] Google proposed the inception module based on low number of parameter. The inception module effectively extracts features from images using a (1 × 1) operation and concatenated the convolution layers in parallel.[3] The AO Foundation proposed classification standard according to the characteristics of the fractures to apply appropriate treatments to a patients. The fractures categories were classified based on the anatomical location, shape and angle of fracture.[4]

3 Characteristics of Multiple fracture Classification Problem

3.1 CT images and target labels

We use CT images of the collected lower body fracture patients for this experiments. The collected data is 105 patient, and each patient has about 24 images. Each image is taken by rotating the CT image of the fracture location at a certain angle as shown in [Fig.1]. However, we use only 80 patient data for experiments because of noise, labeling and other problems of data. The target labels of each data are classified according to "AO / OTA Fracture and Dislocation Classification". The AO Foundation classifies the lower body fractures into a total of 106 categories according to this criterion. In this experiment, we use only 26 labels that appear in the data. The collected data is patients with multiple fractures, each image has between 1 and 4 labels.

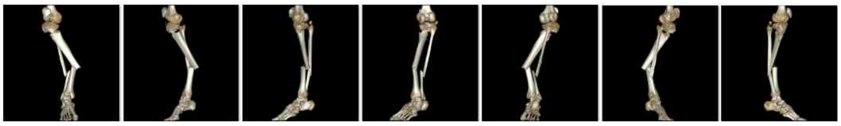


Fig. 1. Example CT image of a patient. In this Figure, each image was taken with a rotation of 45 degrees in fracture location.

3.2 Insufficient and imbalance of data

We use 1897 and 652 images in two experiments because only 80 of the 105 patient data is available. However, the data has an imbalanced problem. The ratio of the four labels among 26 labels is 56%, and the 13 labels is very few. The detailed data distribution based on the number of patients is shown in [Table.1]. We use data augmentation in this experiment to overcome this problem. The data augmentation is applied randomly to each image when training the fracture classification model.

Table 1. Number of data by patient

Label	Num	Label	Num	Label	Num	Label	Num	Label	Num	Label	Num
4F2B	38	4F1B	10	4F3A	7	32B2	3	41A2	1	4F2	1
42B2	37	4F3B	9	4F1A	5	43A2	3	43A1	1	4FB2	1
42A1	21	42A2	8	42C2	4	12A1	2	12B3	1	43B1	1
4F2A	20	42A3	7	43A3	3	43C3	2	31B1	1	32C3	1
42B3	13	32A3	7								

4 CNN Based Multiple Fracture Classification Method

The data used in this experiment is insufficient. For this reason, the model was overfitting in the training data. To prevent this problem, we restructure the model based on the Inception module on GoogleNet to reduce the number of parameters use in the model.[5] [Fig.3] shows proposed model architecture. In back-end of propose model, we connect 26 nodes for the number of labels. We use the sigmoid function as activation function and the sigmoid cross-entropy function as loss function to compute score of each label independently in propose model. That is, the output of the proposed model is the score for each label. After this, the proposed model selects the four labels with the highest scores among 26 label, and applies the threshold to these four label scores. As a result, the fracture classification model has outputs of between 1 and 4 labels.

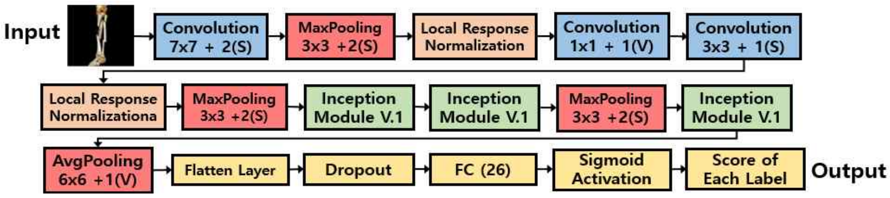


Fig. 3. Proposed model architecture for classification of multiple fractures.

5 Experiment

In this experiment, the precision and the recall are used as metric. The Precision and the recall follow (1) and (2). Here D is number of multi-label data, Y is target label from each data, and Z is predicted label from proposed model.

$$\text{Precision} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|y_i \cap z_i|}{|z_i|} \quad \textbf{(1)} \qquad \text{Recall} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|y_i \cap z_i|}{|y_i|} \quad \textbf{(2)}$$

The data used in the experiment are unevenly distributed by each label. Therefore, we apply K-Fold cross validation [6] to construct the data set, and the threshold is also changed to find the optimum threshold. In the first experiment, the angle of rotation between each image is 15°, 1726 images is used train data and 171 images is used test data. The precision and the recall were the highest at 84.7% and 88.2% when the threshold was 5. However, the limitation of this experiment is how to consist a dataset. The 13 labels among 26 labels were less than 3 patients. For this reason, we consisted a dataset based on a combination of labels in the data. In other words, we used similar images for train datasets and test datasets because the rotation angle of each CT image was 15 °. To improve this limitation, we conducted second experiment and shows that result. In second experiment, we use only a fraction of the total data so that the rotation angle between each images can be 45. This is to reduce the similarity between each image even if each image of the patients was used train dataset and test dataset. We used 547 images for train data and 105 images for test data. The results of the second experiment shows precision was 73.3% and recall was 86.9% when the threshold was 7. Compared to the previous experiment, there is not much difference in the recall. On the other hand, the second precision is 11.4% lower than first precision. That is, the proposed model classified the target label with high accuracy, and not accurately excluded non-target label. However, we assessment that the reliability of the second experiment is higher than the previous experiment through reduced the similarity between the train dataset and the test dataset.

6 Conclusion and Future Work

In this paper, we proposed method to classify fracture based on CT image using deep learning. The proposed method assist the orthopedists identify fractures and further automate that process. We used CT images of the collected lower body fractured patients. We casts the fracture category classification process as a multi-label classification problem, because one CT image has multiple fracture. However, we use only 80 out of 105 patient data due to problems such as noise and etc. For this reason, the data used in experiment has problem such as insufficient and unbalanced. Accordingly, we reduce the number of parameters in the neural network to prevent overfitting. For this, we tune the model based on the Inception module. The proposed model is evaluated the using precision and recall. In the first experiment, precision was 84.7% and recall was 88.2%. However, the first experiment has a limitation such as similar images were used in train dataset and test dataset. To improve this limitation, we set the rotation angle of each image 45° in the second experiment. As a result, we reduced similarity between each image, and the precision was 73.3%, and the recall was 86.9%. Future work is data collect and concentration to overcome limitations from data. In this experiment, we collected data on fractures of all lower body. However, lower body fractures most result in tibia and fibula fractures. [Table.1] shows a high ratio of 4oo label and 4Foo label, especially 42oo, 4F2o. We plan to focus on collect the data of tibia and fibula, and to balance the data besides the 42oo and 4F2o fractures. After that, we will tune proposed model for optimal fracture classification model and extend the classify range of fracture categories. Also, if we collect sufficient data, we should grouping train dataset and test dataset by patient.

Acknowledgments. This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation (NRF) of Korea (Grant no.: NRF-2017M3C4A7069432).

References

1. e-나라지표 : Statistics of Medical Personnel and Sickbed. Available : http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=2772, (Accessed: June 10, 2019)
2. Tsoumakas, Grigorios, and Ioannis K : Multi-Label Classification: An Overview. In: International Journal of Data Warehousing and Mining, pp 1--13, (2007)
3. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., and Rabinovich A. : Going Deeper With Convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1--9, (2015)
4. AO Foundation : AO/OTA Fracture and Dislocation Classification. Available : https://www.aofoundation.org, (Accessed: June 10, 2019)
5. Dziugaite, Gintare K., and Daniel M. : Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data, In: ArXiv Preprint, ArXiv:1703.11008(2017)
6. Bengio Y., and Grandvalet Y. : No Unbiased Estimator of the Variance of K-fold Cross-Validation, In: Journal of Machine Learning Research 5(Sep), pp 1089--1105, (2004)

Activity-based Software Maintenance Cost Estimation for Package Software in Korea

Kyoung-ae Jang [1,1], Woo-Je Kim [1,2]

[1] *Dept. of Industry and Information Systems Engineering, Seoul National Univ. of Science and Technology,*
232, Gongneung-ro, Nowon-gu, Seoul, 01811, Republic of Korea
{ Frontier Laboratory Room 701}

Abstract. This paper defines software maintenance activities and proposes a model for software maintenance cost estimation of package software. First, we review the literature surveys to provide software maintenance activity classifications or activity factors. Then, we develop a cost structure for package software based on the identified activity factors, and propose an activity-based cost estimation model. Finally, we verify the model with actual cost data from software maintenance projects.

Keywords: software maintenance, cost estimation, cost model, package software, cost activity

1 Introduction

The proportion of software in systems is increasing, and software has become an indispensable factor. A 2010 survey showed that approximately17.6% of software sales in package software companies in Korea were related to maintenance of package software [1]. Revenue from package software consists of license costs and maintenance costs. The cost of licensing the product is dependent on the maintenance cost policy. The software maintenance cost is an important factor in the package software revenue model.

Although the package software sector is growing, there are insufficient standards in the software maintenance market of the sector. It is difficult to standardize software maintenance activities because there exist several business activities that are related, and there is no clear demarcation between defect repair, maintenance, and free servicing activities. We defined software maintenance activity elements for package software and developed an activity-based software maintenance cost estimation model.

The rest of this paper is structured as follows: Section 2 introduces the theoretical background for this study. Section 3 discusses design of the model and verifies the developed model. Section 4 provides the conclusions and discusses future research.

[1] First Author

[2] Corresponding Author

2 Classification of software maintenance activities

We categorized maintenance activities through analytical research. Previous studies included components of ISO / IEC 14764 and IEEE 1219, maintenance activity elements published by the International Function Point User Group (IFFUG), the United Kingdom Software Maintenance Association (UKSMA) and the Korea Software Industry Association (KOSA) [2][3][4][5][6][7][8][9].

Through the results, we integrated the activities based on the guideline and classified the activities into product service, technology service, and support service using the KJ technique. Some of the criteria we integrated into those activities areas are as follows:

- Pattern maintenance includes error correction and version patch of the package software.
- Certification includes services to quantify how well the package software meets the quality assurance plan.
- Improvement services includes adaptive and corrective patch services, and developing new defenses for hacking
- Technical services include disaster recovery, help desk support, and preventative maintenance.

We developed a standard model for software maintenance activities based on the above criteria and the previous studies. The model has product service, technology service, and support service. The product service consists of an adaptive patch service and a corrective patch service. The technology service consists of online support, emergency visit service, and regular visit service. The support service consists of training for operation and quality assurance and certification.

3 Software maintenance cost estimation model

The software maintenance costs are made up of the direct labor cost, indirect overhead expense, engineering fee, direct overhead cost, and value added tax (VAT). According to the engineering cost estimation model, the overhead expense can be estimated as110%–120% of the direct labor cost and the engineering fee can be estimated as 20%–40% of the sum of the direct labor cost and the indirect overhead expense [6, 10]. To calculate the direct labor cost, the number of occurrences, the average service time required, the number of workers required, and the manpower level for each software maintenance activity are obtained. As this result, we proposed software maintenance cost structure model for package software.

Then we have furthermore studied the method of applying the adjustment factors to lower the estimation error rate and raise the accuracy in the developed software maintenance cost estimation model.

To validate the model, we collected the cost data and data for adjustment factors for 19 actual package software maintenance projects from domestic companies, and compared the estimated cost obtained by the proposed model with the actual contract price for each project. Three measurements were used to verify the model: the coefficient of determination (R^2), magnitude of relative error (MRE), and prediction quality (PRED) measure. The performance of the developed model was very high (R^2> 99%, MMRE = 3%, and PRED(0.25) = 94%).

4 Conclusion

This study investigated package software maintenance activities by reviewing previous research, classifying the identified activity elements, and proposed a cost estimation model based on the software maintenance activities. The activity-based model calculates the direct labor costs and estimates the total cost for the software maintenance. In addition, we developed the adjustment factors to improve the accuracy of the model.

To validate the model, we collected the cost data and data for adjustment factors for 19 actual package software maintenance projects. The performance of the developed model was very high.

This shows that the proposed model can be reliably applied to predict the package software maintenance costs for actual projects. It is necessary to continuously collection and verification the actual cost data and adjustment factor data to verify various perspectives.

References

[1] W. J. Kim and S. J. Jung, A Method for Estimating Maintenance Cost of Package Software, in Proc. of 4. International Conference on Computer Science and Information Systems, Dubai, UAE, 2014, 43–46

[2] International Standard ISO/IEC 14764, Software Engineering-Software Life Cycle Processes-Maintenance (ISO/IEC), 2006

[3] IEEE 1219, IEEE Standard for Software Maintenance (IEEE), 1998

[4] IFPUG,The IFPUG Guide to IT and Software Measurement ,CRC Press - Taylor & Francis Group, FL, USA, Boca Raton, 2012

[5] ISBSG, Managing Your Maintenance & Support Environment 2012 Update ,ISBSG, Melbourne, Australia, South Melbourne, 2012

[6] UKSMA, Measuring Software Maintenance and Support, Version 0.5, Draft,http://www.uksma.co.uk/,July 1st, 2001

[7] Korea Internet and Security Agency, A guide for cost estimation of software project (2016 edition) ,Korea Internet and Security Agency, Korea, 2016

[8] Korea Internet and Security Agency, A guideline for maintenance service of open source software, Korea Internet and Security Agency, Korea, 2007

[9] Korea Ministry of Information and Communication, Package software maintenance service guidelines, Korea Ministry of Information and Communication, 2005

[10] Korea Ministry of Knowledge and Economy, A Guide for Cost Estimation of Engineering Projects ,Korea Ministry of Knowledge and Economy, Notification 2012-178, 2012

J. Bloem, M. Van Doorn, S. Duivestein, D. Excoffier, "The Fourth Industrial Revolution", sogeti.com., 2014.

A Data Set Management Approach to Mitigate Data Aging

Tae-Hyung Kim[1] and Seo-Young Noh[2*],

[1] Samsung Electronics, 56 Seongchon-gil, Seocho-Gu, 06765, Seoul, Korea
thkim4u@hanmail.net
[2*] Department of Computer Science, Chungbuk National University, Chungdae-ro 1, Seowon-Gu, 28644, Cheongju, Korea,
rsyoung@cbnu.ac.kr

Abstract. Big data is usually described with multiple V's to represent its characteristics in industrial domains and academic researches. Big data is a set of huge, complex and unstructured data. It is practically difficult for a traditional system and database to process effectively. In this paper, a generic framework for processing big data and its relationship with those characteristics of big data are discussed. To reserve and maintain the specific data sets used to generate information via the big data processing framework, the conceptual methods for managing those data sets are proposed in order to keep pace with the fast growth and frequent change of big data.

Keywords: Big Data Characteristics, Big Data Processing Framework, Software Engineering, Data Aging, Data Recalling, Data Reassessing

1 Introduction

Big data in a broad sense refers to a research area for treating very large and compound data sets. It is ordinarily characterized by four to ten words beginning with 'V' [1, 2, 4, 6]. Among them, volume, velocity, variety and veracity are considered as the primary characteristics. Volume comes from the tremendously large size of big data. Velocity is the changing speed of big data that is produced considerably faster, often in real-time. Variety represents the forms of big data of which type is unstructured and unorganized. Various types of big data are roughly classified in [5]. Veracity means uncertainty of big data, so it gives a question of reliability of data source and confidence in data itself during analyzing. In addition to the four V's that denote the main properties of big data, variability refers to consistency of big data or describes the multitude dimensions appearing during big data analysis. The most important thing that should be delivered to stakeholder is value to be excavated from big data. For this purpose, the big data processing framework is presented and the relationship of its internal stages with the characteristics mentioned above are discussed in the following section.

Those V characteristics could be regarded as environmental or operational issues. Actually, they are related to various quality attributes, such as availability, security, performance, usability, as well as functional requirements that are satisfied during the development of real big data systems. The big data system design method (BDD), as a combined process model of architecture design with data modeling, is proposed to help design and develop them [2].

[*] Corresponding Author

2 Big Data Processing Framework

The main objective of processing big data is to *mine* a scientific or business value hidden in big data. It is not so easy for big data to be treated through a simple step and analyzed with a traditional software solution. Therefore, a big data processing system is equipped with advanced technologies and AI-based methods to capture, transform and analyze enormous data sets. Fig. 1 shows the representative framework with six stages required for processing big data. The five stages, except the last integration stage, in this big data processing framework correspond closely with the process steps for big data analytics application development that are acquisition, presentation, preprocessing, processing and generating/presenting results [4]. Also, the process for discovering knowledge in databases, called KDD process, has analogous steps [7].

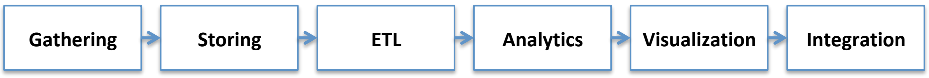


Fig. 1. A big data processing framework with six stages

First, the gathering stage identifies where to find raw data and how to crawl it. Second, some of the raw data is moved to the repository with a simple structure like a modern NoSQL database. Third, a set of the collected data is transferred to a database or data warehouse during the ETL (Extract-Transform-Load) stage. Fourth, the analytics stage makes a deliverable that can explain a current situation, provide insight or predict the future trends. The deliverable contains abstract models, meaningful results, and so on. Statistical approaches and AI techniques including machine learning and natural language processing are aggressively applied during this stage. Fifth, the visualization stage harnesses the deliverable produced after big data analysis to help stakeholders see the information extracted from big data. Most of big data systems generally cover from the storing stage to the analytics or visualization stage. They are implemented using either open source solutions such as Hadoop, Spark and Strom or proprietary solutions such as Google File System and Amazon DynamoDB [1]. The lambda architecture is introduced especially when a data processing system is designed to handle massive data by means of both batch and stream mode.

Finally, the integration stages is performed only if it is necessary for the visualized information or portions of the deliverable to be integrated into a working system as a form of a source code, a concrete service or a microservice. As this stage follows either a traditional software development process or an agile software development method like Scrum [4], the implementable features that are fundamental requirements units should be identified based on the information and deliverable obtained from the previous stages. If these features are changed frequently due to some of the V characteristics shown in the previous section, this integration stage may require extremely high maintenance effort. For example, while a machine learning code is at most 5% in a mature system, the rest of 95% is clue codes [3], which causes high coupling with other modules and increases maintenance cost as well as technical debt.

Table 1 shows that each stage in Fig. 1 is mostly relevant to, but not limited to, one of V characteristics. For instance, the visualization stage highly affected by variability could be difficult to denote some information with high accuracy and completeness because of veracity.

Table 1. Each stage is mapped to a mostly related characteristic of big data.

Stage	Input	Output	Mostly related to
Gathering	Data sources	Raw data	Velocity
Storing	Simple structure	Data collection	Volume
ETL	Rules	Refined data	Variety
Analytics	Hypothesis	Models or Results	Veracity
Visualization	Deliverables	Information	Variability
Integration	Features	Services or APIs	Value

3 Data Management Methods

Software aging [8] occurs due to failure of meeting new requirements or catching the change of technologies and trends, which decreases the value of software and its quality attributes like maintainability or understandability. Software aging may be detected because of its malfunction or increased maintenance cost. Software restructuring improves software quality by transforming its internal structure and keeping its external interfaces to users equivalent or compatible. In particular, code refactoring reduces the internal complexity of software developed using object-oriented methods without changing its behaviors. Software restructuring and code refactoring are expected to make software aging slow down, but not stopped. Data aging, however, is unpredictable and inescapable. For example, it is almost impossible to guarantee that all of the shopping sites used as data sources will be available in the next month or year.

Data processing cost and time tend to be directly proportional to the size of big data treated and strongly dependent on the infrastructure applied. Even in big data era, it is meaningful to reserve and track critical data sets used to generate information for current and future needs. The critical data set that contributes to the particular deliverable is called the principal data set. In turn, a deliverable is supposed to be a kind of convergence point from which its corresponding principal data set could be identified and segregated by tracing backward to the big data repository. A principal data set could be the same as all of the big data that is currently stored in a repository and entirely exploited, for instance, by a machine learning approach.

To mitigate data aging by coping with the rapid change of big data, a principal data set needs to be periodically examined and carefully maintained. It is possible for some obsolete data to degrade the value or quality of a deliverable. Accordingly, the service could not work as intended when its features are related to that deliverable. In order to manage the principal data set selected from the big data repository and used to produce a deliverable, we propose the data management methods deduced from the software restructuring and code refactoring. Fig. 2 describes the conceptual overview of the two methods: data recalling and data reassessing.

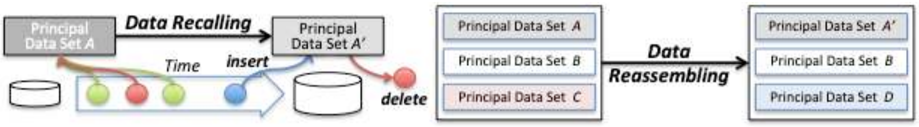


Fig. 2. The conceptual overview of the data recalling and the data reassessing

The data recalling is the method of reconstituting a principal data set by inserting new data and removing unnecessary, unused or outdated data. The principal data set after recalling should not take any effect on deliverable's value and quality, which could be measured with predefined metrics such as accuracy or precision. For the deliverable that is dependent on multiple principle data sets, the data reassessing is proposed as the process of unifying several principal data sets, replacing some principal data sets, removing the data duplicated across multiple principle data sets and identifying a new one from the principal data sets initially given. This process can include the data recalling for an individual principal data set if necessary. The result of data reassessing is evaluated by scrutinizing whether the principal data sets are well categorized or classified, so that the number of the principal data sets before and after reassessing is irrelevant.

4 Conclusion and Future work

As the major characteristics of big data, the six words starting with 'V' are addressed. The big data processing framework is designed to deal with very large, complex and unstructured data with consideration of those V characteristics. To manage the principal data set directly contributing to the deliverable obtained as a result of processing big data, data recalling and data reassessing are proposed as data management methods. The main purpose of these two methods is to retain the principal data sets against data aging.

Currently we do not have a well-established method to determine a principal data set. Also, two methods need to be formally described and procedurally explained with detailed steps. Hence, the next step of our research focuses on application of our proposed methods with practical case studies. The recommendation and automation tools could be another direction of future work.

Acknowledgments. This work was supported by the research grant of the Chungbuk National University in 2019.

References

1. Karakaya, Z.: Software Engineering Issues in Big Data Application Development. In: IEEE 2nd International Conference on Computer Science and Engineering, 851--855 (2017)
2. Chen, H., Kazman, R., Haziyev, S., Hrytsay, O. : Big Data System Development: An Embedded Cast Study with a Global Outsource Firm. In: IEEE/ACM 1st International Workshop on Big Data Software Engineering, 44--49 (2015)
3. Sculley, D., Holt, G., Golovinm D., Davydov, E., Phillips, T., Enber, D., Chaudgary, V., Young, M.: Machine Learning: The High Interest Credit Card of Technical Debt. Software Engineering for Machine Learning (NIPS 2014) Workshop (2014)
4. Al-Jaroodi, J., Hollein, B., Mohamed, N.: Applying Software Engineering Process for Big Data Analytics Application Development. In: IEEE 7th Annual Computing and Communication Workshop and Conference (2017)
5. Lv., Z., Song, H., Basanta-Val, P., Steed, A., Jo, M.: Next-Generation Big Data Analysis: State of the Art, Challenges, and Future Research Topics. In: IEEE Transaction on Industrial Informatics, vol. 3, no. 4, 1891--1899 (2017)
6. Madhavji, N., Miranskyy, A., Kontogiannis, K.: Big Picture of Big Data Software Engineering. In: IEEE/ACM 1st International Workshop on Big Data Software Engineering, 11--14 (2015)
7. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Data Bases. In AI Magain, vol. 17, no. 3, Fall 1996.
8. Parnas, D.: Software Aging. In: ICSE '94 Proceedings of the 16th International conference on Software Engineering, 279—287 (1994)

A Design of Data-based City Policy Knowledge Model

Sun-Young Ihm[1], So-Hyun Park[1] and Young-Ho Park[1,*]

[1] IT Engineering, Sookmyung Women's University, Seoul, Korea
{sunnyihm, shpark, yhpark}@sm.ac.kr
*corresponding author

Abstract. Use of data as evidence has a significant role in public policy development. Data-based policy decisions are based on rational analysis and can provide credibility by providing justification for policy. In this paper, we design a policy knowledge structure model for policy development. In addition we propose a utilizing service with the proposed policy knowledge model.

Keywords: Policy development, Policy knowledge model, Data-based policy

1 Introduction

Recently, the importance of policy decision based on objective data has been emphasized rather than policy based on human intuition, especially in developed countries such as the US and Europe. For example, In the United States, the policies were established to prevent tax evasion and manage energy in buildings based on data. Also in Japan, disaster response policy was established based on data, and in Korea, bus route policy was established based on floating population and mobile phone call data[1-3].

Data-based policy is widely understood as evidence-based policy use derived from data, research results, and policy evaluations[4]. There are a lot of information to consider for the establishment of city policy, but there are few studies that systematically established it. In addition, most studies have been developed only for specific city situations and policies in the form of case by case, so there is a problem to be newly developed for other policies.

In this paper, we design a policy knowledge model for city policy establishment. A reliable policies can be developed and social/economic costs can be reduced by using the proposed policy knowledge structure model. In addition, knowledge-related queries and reasoning are possible. The rest of this paper is organized as follows. Section 2 describes the proposed policy knowledge structure model and Section 3 concludes the paper.

2 Policy Knowledge Model

Opinion-based policy decisions are based on unreliable perspectives such as individual opinions, biases, and conjectures, but data-based policy decisions are based on rational analysis and can provide credibility by providing justification for policy. The figure 1 shows the proposed policy knowledge model. In the first step, we would collect various data such as geographic data, sensor data, and social network service data. Next, the policy knowledge is extracted using various methods such as statistical analysis, simulation, data mining, and machine learning. Finally, we would implement a service that manages and utilizes the policy knowledge structure.

Since the policy knowledge is structured, it is possible to query for policy development. For example, the answer can be derived as a related factor to the query, “Extract influential factors to activate city business district”.

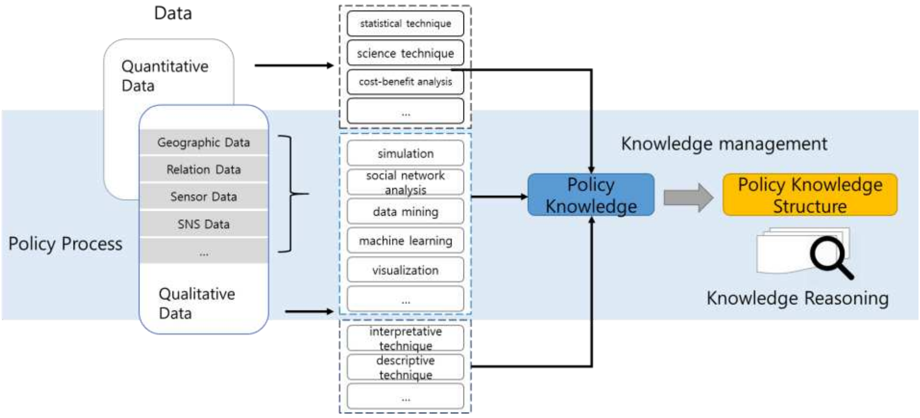


Fig. 1. A proposed policy knowledge model.

4 Conclusion

In this paper, we design and propose the policy knowledge model for data-based policy development. The policy knowledge structure would be extracted by analyzing data, and the policy factor could be derived from the policy knowledge structure. As for the future work, we would implement the policy knowledge system and reasoning service.

Acknowledgments. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2016-0-00406, SIAT CCTV Cloud Platform)

References

1. M. Tenney, R. Sieber.: Data-Driven Participation: Algorithms, Cities, Citiznes, and Corporate Control. Urban Planning, vol.1, no.2 (2016)
2. S. Verhulst, J, Z. Engin, J. Crowcroft.: Data & Policy: A new venue to study and explore policy-data interaction. Data & Policy, vol.1 (2019)
3. S. Kim, K.S, Chung.: Policy Implementation and Alternative Policy Options of Big Data: Issues of Present Condition and Policy Application with Big Data, Korean Comparative Government Review, vol.18, no.3, pp.309-324 (2014)
4. K.S. Yoon.: Enhancing the Use of Data as Evidence in Public Policy Development. Technical Report, Korea Institute of Public Administration (2016)

Software Component Reuse using Skyline Query

Jong-Hyeok Choi[1], Aziz Nasridinov[1,*]

[1] Department of Computer Science, Chungbuk National University, 28644, Cheongju, Korea
{leopard, aziz}@chungbuk.ac.kr, *Corresponding author

Abstract. Recently, software companies have been reusing pre-verified software components to ensure product reliability and shorten development time. However, it is challenging to select optimal components in the software design stage. To solve this problem, we propose the theoretical background of the component selection method using the skyline query, a multi-criteria decision-making technique, in order to select optimal components and their combinations.

Keywords: Software reuse, Skyline query, Multi-criteria decision-making

1 Introduction

Recently, companies have been accelerating the development of software by reusing software components [1-2]. This change in the software development environment occurs because of various components are provided to developers through open source projects and commercialized components. These pre-verified components can provide rapid development and high reliability in the software design stage. However, it is challenging to select optimal components and their combinations from components that perform similar or same functions. Also, considering that the choice of the wrong component leads to an increase in time and cost of software development, optimal component selection is a significant problem.

To solve this problem, we propose a new method to select optimal components and their combinations in the software design stage using the skyline query, a multi-criteria decision-making techniques [3]. In this paper, we present a theoretical background of the proposed method that contains three main phases: combination graph modeling, component domination query, and optimal component selection.

2 Component-based Software Reuse using Skyline Query

In this section, we describe three phases of the proposed method: (1) *combination graph modeling*, (2) *component dominant query,* and (3) *optimal component selection*.

Combination graph modeling. The combination graph modeling performs component modeling by arranging combinable components in functional units. The model contains input layer, functional layer, and output layer. The functions of the software are divided into functional layers and sequentially stacked according to the processing order. Also, each layer declares a component that performs a necessary function or input data as a node and generates an edge when a combination between nodes is possible.

Component domination query. The component domination query selects optimal components from each functional layer using various performance metrics with the skyline query and removes unnecessary components from the model. For this, the component domination query decomposes the performance metrics like equations as shown in Table 1, maps them as points in the multi-dimensional space, and then compares the values of the same dimension through the skyline query. However, when components have different performance metrics, they do not have a specific dimension value. To solve this problem, the proposed method makes it possible to compare using uncertain skyline query [3]. This query can reduce the number of combination that occur between components belonging to different functional layers by removing components that always lower performance than others.

Optimal component selection. The optimal component selection selects the best component combination using the components that are not dominated through the component domination query and the input values of the operating environment expected at the design stage. In this phase, the expected time, resources, and others are input values to the nodes. The input layer is calculated through layer order and performance metrics, and the final component combination is selected based on these result. Considering that the selected combination of components is selected based on the performance metrics represented by numerical values, it makes it possible to select objective components rather than a software designer's perception.

Table 1. Multidimensional Data Space Mapping Results of Performance Metrics

Comp.	**Metrics**		**Dimensions**						
	Time	Memory	x	y	z	yz	y^2	y^z	n
A	4x+2y	n	4	2	*	*	*	*	1
B	4x+2y+3z	n	4	2	3	*	*	*	1
C	2x + yz	2n	2	*	*	1	*	*	2
D	2x + 6yz	3n	2	*	*	6	*	*	3
E	$x + y^2+3z$	4n	1	*	3	*	1	*	4
F	$x + y^3+3z$	4n	1	*	3	*	y	*	4
G	$x + y^z$	2n	1	*	*	*	*	1	2

3 Conclusion

In this paper, we have proposed a new method to select optimal component and combination of them through skyline query in order to effectively reuse software components. In future research, we will develop the proposed method as a practical tool and apply it to the development environment that performs component reuse.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3B03035729).

References

1. Ichii, M., Hayase, Y., Yokomori, R., Yamamoto, T., Inoue, K.: Software component recommendation using collaborative filtering. In: ICSE Workshop on Search-Driven Development-Users, Infrastructure, Tools and Evaluation, pp. 17-20. IEEE. Vancouver (2009)
2. Pande, J., Garcia, C. J., Pant, D.: Optimal component selection for component based software development using pliability metric. SIGSOFT Software Engineering Notes. 38(1), 1-6 (2013)
3. Gulzar, Y., Alwan, A. A., Abdullah, R. M., Xin, Q., Swidan, M. B.: SCSA: Evaluating skyline queries in incomplete data. Applied Intelligence. 49(5), 1636-1657 (2019)

A Novel on Error Correction Model of Electronic Power Big Data based on Missing Value Imputation using MLE

Se-Hoon Jung*, Jun-Ho Huh**
*Assistant Professor of **School of Connection Major (Bigdata Convergence), Youngsan University of Yangsan**, Republic of Korea
*Email: shjung@ysu.ac.kr
Assistant Professor of Dept. of Software, **Catholic University of Pusan, Republic of Korea
**Corresponding Author Email: 72networks@cup.ac.kr or 72networks@pukyong.ac.kr

Abstract. The data collected for the power transmission towers built in an open area is often largely influenced by the noise or deflection caused by external forces so that it has a characteristic of making it harder to determine an abnormal situation. In order to set a threshold to make such a judgment, it is essential to secure a stable data, especially the raw data that does not include any missing data or outliers. The recent large-scale forest fire in Ganwondo province of the Republic of Korea was caused by the embers from a transmission power constructed in open area and such a disaster could have been prevented if there was an accurate diagnostic system along with a rapid response system. However, the sensor equipment installed on the existing transmission towers often cause data omission or missing as well outliers, resulting in the low-accuracy analysis. This study proposes two types of algorithms and develops an analysis model by integrating them. Different from the existing replacement algorithms, the first algorithm is a correction algorithm in which maximum-likelihood estimation (MLE) and K-Nearest Neighbor (k-NN) algorithm have been combined to replace the missing data with all sample data. The second algorithm was designed by combining a modified K-means algorithm with a principal component analysis algorithm to check abnormalities or outliers in the power data based on the corrected raw data. Then, an analysis and prediction model for the power data of power transmission towers were developed by combining these algorithms. The proposed k-NN+MLE algorithm showed better results compared to the existing missing value replacement algorithms across all the data having missing values and the analysis and forecasting model applied wit a modified K-means algorithm showed the improved results with the increased accuracy from 3.064% to 0.366% compared to the existing power data analysis model.

Keywords: Electronic Power Big Data; Big Data; k-NN; MLE; Python

1. Introduction

The importance of power data is increasing every day across all the industries in the midst of the 4th Industrial Revolution as it is possible to control or distribute the power stably by analyzing and forecasting it [1-2]. The power data analysis and forecasting system is an essential research area since human or material damages are increasing due to the increase in electricity or power-related incidents [3-5]. Especially, the increase in electrical energy consumption by various types of industries all across the globe shows the absolute necessity of stable power supply and accurate analysis of power systems [6-7]. That is, an accurate and stable power analysis system is absolutely required from the supply side whereas an analysis management system is an absolute requirement from the demand side.

The power transmission tower installed in an open-air location and to be used for this study is a part of a standard power facility which can measure the power data. Some of the small damages caused to a power transmission tower by external environment or natural disasters (external forces) could lead to much larger damage such as the large-scale forest fire which had occurred in Gangwondo province in 2019 [8-12]. Like such an incident, the damages to the transmission towers largely affect both the economy and society so that it is essential to maintain them appropriately to minimize such negative effects [13-17]. Currently, the Korean government is servicing transmission towers through Korean Electric Power Corporation (KEPCO) who continuously collects data with the sensor nodes installed on them. However, such an attempt is sufficient enough to detect the outliers occurring due to the limited measuring range of the sensor or the noises caused by natural phenomena or external forces.

An algorithm in which a modified Kmeans and a principal component analysis are being combined is proposed to identify the problems in the power data or occurring due to external forces.

2. Novel on Analysis and Prediction Model of Transmission Line Tower Big Data

Fig. 1 is showing the structure of the analysis & prediction model proposed in this study, which is largely divided into two parts: Analysis and Prediction and Missing Value Imputation. The former performs a preprocessing based on the calibrated data and the preprocessing module includes removal of unnecessary data in the power data as well as its normalization to perform a power data analysis and prediction operation by combining a modified K-means algorithm and a principal component algorithm. The latter performs missing value imputation through the k-NN + MLE calibration algorithm.

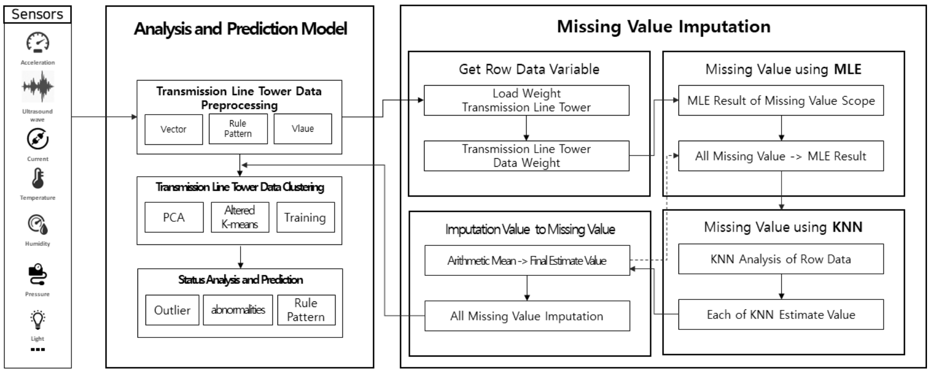

Sensors
Acceleration
Ultrasound wave
Current
Temperature
Humidity
Pressure
Light
Analysis and Prediction Model
Transmission Line Tower Data Preprocessing
Vector
Rule Pattern
Vlaue
Transmission Line Tower Data Clustering
PCA
Altered K-means
Training
Status Analysis and Prediction
Outlier
abnormalities
Rule Pattern
Missing Value Imputation
Get Row Data Variable
Load Weight Transmission Line Tower
Transmission Line Tower Data Weight
Missing Value using MLE
MLE Result of Missing Value Scope
All Missing Value -> MLE Result
Imputation Value to Missing Value
Arithmetic Mean -> Final Estimate Value
All Missing Value Imputation
Missing Value using KNN
KNN Analysis of Row Data
Each of KNN Estimate Value

Fig. 1. Structure Diagram of Proposed Model

Power will be decreased due to the reduced number of samples when the analysis has been performed by ignoring the missing values in power data and such a problem can be avoided by imputing them with the relevant patterns or expected values rather than simply discarding them. The k-NN algorithms were used in the existing studies for missing value imputation or the merit of not requiring any hypothesis for the data distribution, different from statistical approaches. However, when there is no consistent or similar pattern, the effectiveness of missing value imputation by these algorithms can be reduced. As the size of power data can be large, small, or has no consistent pattern, research for an imputation method considering the data size has been required. This study proposes a k-NN + MLE algorithm to deal with such these problems. MLE is a statistical pattern estimation technique which finds an optimal θ and it is possible to calculate an estimated value at which the probability of an event becomes maximum. The accuracy of this technique becomes higher as the data size increases. As shown in Fig. 2, there are five patterns in the collected power data: There was a normal data in section A (partial existence of the missing values or outliers), B, C (partial existence of the missing values or outliers), D, and E (partial existence of the missing values or outliers) but in the two sections M _{1} and M _{2} represented in blue, there were missing values, to which mean, conditional, and multiple imputations were used originally.

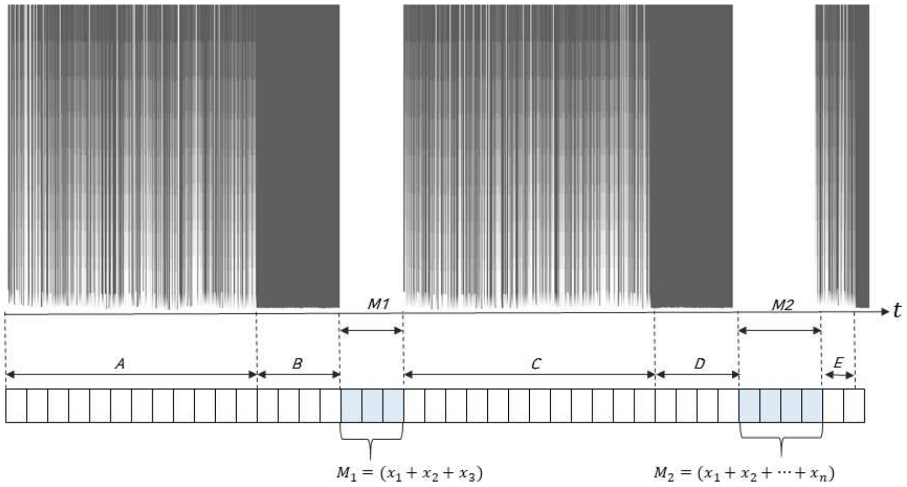


Fig. 2. Example of Electric Data in Missing Value

However, when the K-NN algorithm is applied, there will be a problem of not being able to provide effective observation information for the imputation measurement sections A, C, and E where there are patterns including missing data values or effective data when performing imputation despite the fact that they are not the sections in which no missing values have been generated. As a result, the imputation will be performed for the sections M_{1} and M _{2} by using sections B and D only [Fig. 3].

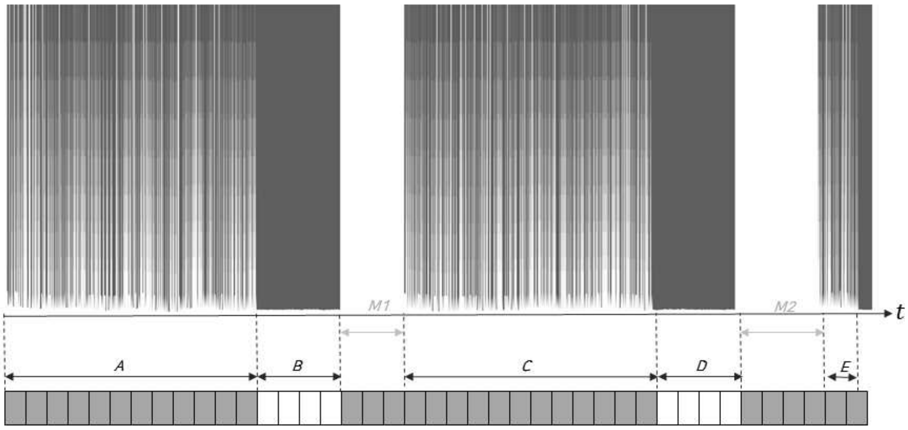
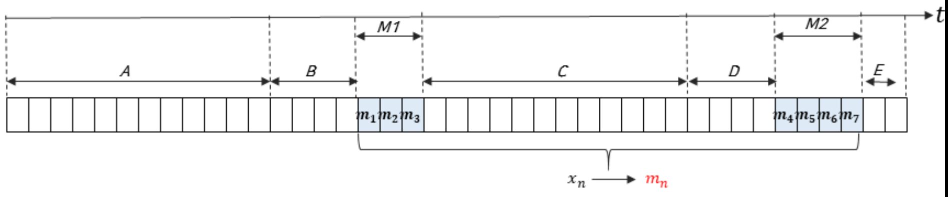
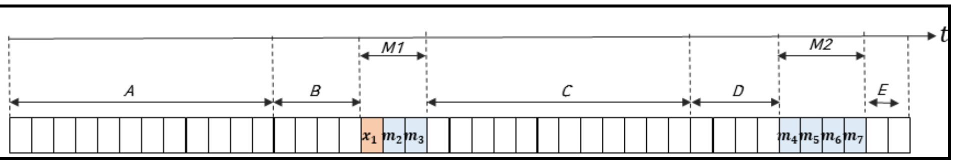
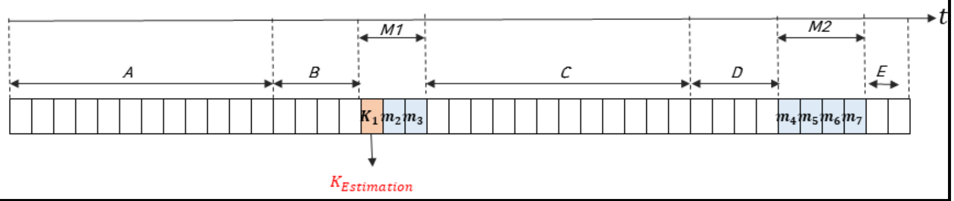
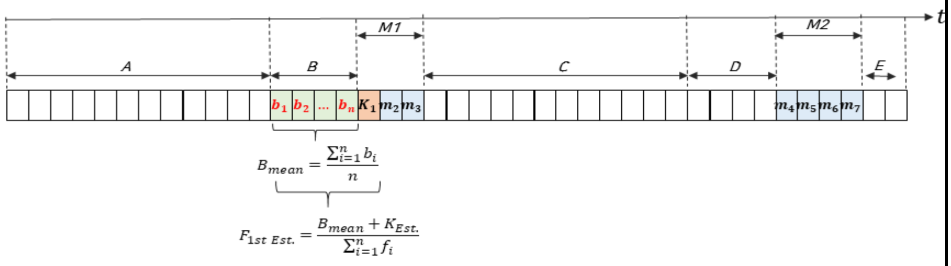
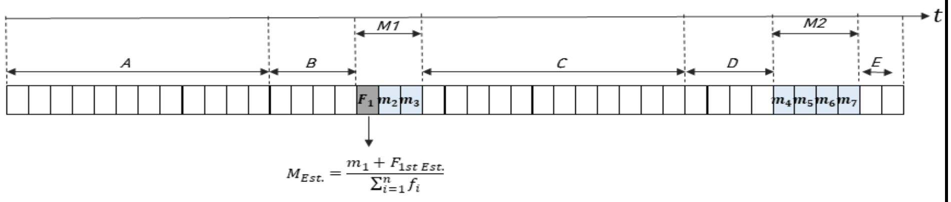


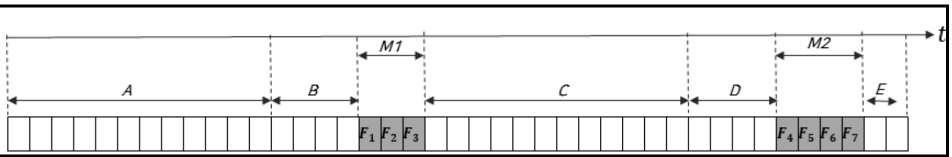
Fig. 3. Example of Electric Data in Missing Value (K-Nearest Neighbor Algorithm)

This study proposes a method of imputing missing values by using the sections A, C, and E which have not been used, as shown in Fig. 3. For this, MLE is used to solve the problem mentioned above. MLE uses a mechanism which finds an optimal θby utilizing entire data: All the sections in the power data except the sections where data is missing are used to calculate the estimated values for the imputation of missing values. The k-NN + MLE missing value imputation algorithm for the power data analysis and prediction is shown in Fig. 4 and its step-by-step processing methodology is shown in Tab. 1.

Table.1. Step-by-Step Processing Methodology

Step	Description
Step 1	Temporarily change all the missing values in the power data with the values resulting from MLE.
	 t M1 M2 A B C D E m_1 m_2 m_3 m_4 m_5 m_6 m_7 $x_n \longrightarrow m_n$
Step 2	Among the instances which have been changed with the resulting values obtained from MLE, change only one instance as a missing value and impute the rest with the MLE values.

	 t M1 M2 A B C D E x_1 m_2 m_3 m_4 m_5 m_6 m_7
Step 3	Calculate the estimated value for the instance having a missing value by applying the k-NN estimation.
	 t M1 M2 A B C D E K_1 m_2 m_3 m_4 m_5 m_6 m_7 $K_{Estimation}$
Step 4	Calculate the arithmetic mean by combining the mean value of the prior power data with the k-NN estimated value as a primary estimation.
	 t M1 M2 A B C D E b_1 b_2 ... b_n K_1 m_2 m_3 m_4 m_5 m_6 m_7 $B_{mean} = \frac{\sum_{i=1}^{n} b_i}{n}$ $F_{1st\ Est.} = \frac{B_{mean} + K_{Est.}}{\sum_{i=1}^{n} f_i}$
Step 5	Calculate the arithmetic means for the MLE value and the primary value to determine the final estimation.
	 t M1 M2 A B C D E F_1 m_2 m_3 m_4 m_5 m_6 m_7 $M_{Est.} = \frac{m_1 + F_{1st\ Est.}}{\sum_{i=1}^{n} f_i}$
Step 6	Repeat Step 1 to Step 5 until all the estimations for the missing values have been calculated.
Step 7	Generate a complete missing value data vector after storing the final estimated values.

<table>
<tr><td></td><td>
</td></tr>
<tr><td colspan="2">x _{n} : data missing instance, m _{n} : MLE estimation, K _{n} : k-NN estimation, f _{n} : primary missing data estimation, F _{n} : final missing data imputation value</td></tr>
</table>

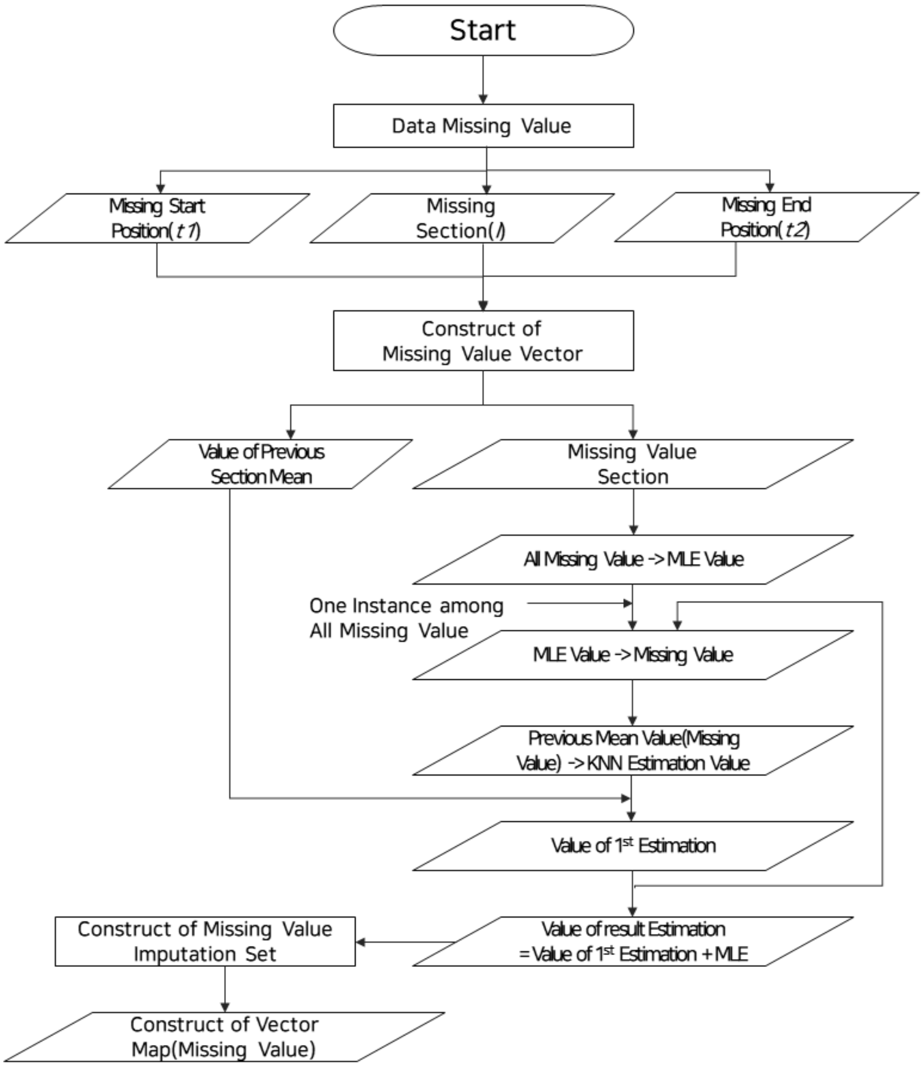

Start
Data Missing Value
Missing Start Position(t1)
Missing Section(l)
Missing End Position(t2)
Construct of Missing Value Vector
Value of Previous Section Mean
Missing Value Section
All Missing Value -> MLE Value
One Instance among All Missing Value
MLE Value -> Missing Value
Previous Mean Value(Missing Value) -> KNN Estimation Value
Value of 1st Estimation
Value of result Estimation = Value of 1st Estimation + MLE
Construct of Missing Value Imputation Set
Construct of Vector Map(Missing Value)

Fig. 4. Flow chart of Proposed k-NN MLE Method

First, the proposed algorithm imputes the missing values included in the power data with the values obtained from MLE performed against the entire data collected. This is to allow utilization of the observation data of instances including missing values when the k-NN algorithm is applied. Second, among the instances where all the missing data values have been initialized with the values resulting from MLE, only one is selected to change its MLE value into a missing value. This is to apply the k-NN algorithm to entire data except for the changed instance to calculate an estimated value for the missing value. Third, for the instance having a missing value, an estimated value is calculated by performing the imputation by applying the k-NN algorithm. In this process, the entire observation data of the instances containing missing values can be used as all the instances which have been imputed initially will be used for the calculation using the k-NN algorithm. In the existing k-NN algorithm, the section M1 _{1} has been excluded when calculating the imputed value in the section M1 _{2} but the proposed algorithm can calculate this value by using the MLE value applied in the first step as an observation data after storing the missing value of section M1 _{1} in a temporary vector table. This is the contribution of the proposed algorithm.

In the fourth step, a primary estimation is calculated for the instance having a missing value by calculating the arithmetic mean of the prior power data section where data is missing together with the arithmetic mean of the k-NN estimated value calculated in the third step. This process is added to reflect the perpetuity and continuity of the data collected earlier considering the characteristics of power data. In the fifth step, the final estimation is calculated by finding the arithmetic mean of the MLE value of each instance computed for the missing value that has been changed in the first step and the primary estimations calculated in the fourth step. This process is repeated until all the instances are satisfied and, finally, after these final estimations are stored to create a complete vector in which missing values have been imputed with the estimations, missing data imputation will be completed. Then, in order for this final estimation to avoid affecting the other estimations, it will not be reflected in the data immediately but stored in the final result temporarily.

3. Conclusion

The transmission towers and power structures are quite sensitive to environmental factors or external forces which make it difficult to perform adequate maintenance and keep in good condition. Also, the existing power structures do not guarantee complete securement of IoT-related data without any loss. In this study, a real-time analysis and prediction model for the power structure has been proposed to reduce damages or avoid any harmful incidents. The data collected by the power structures installed in the past often have many outliers or missing data and there have been a number of problems in the methods of analyzing or predicting such data. Thus, research on an analysis and prediction model which allows identification of outliers or abnormality in the data followed by effective and rapid imputation of missing values was proposed. This model is expected to efficiently analyze the problems in the power data based on the calibrated raw data. As a result, a system which can adequately utilize the data obtained from the transmission towers was constructed by using the k-NN + MLE algorithm was introduced.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2017R1C1B5077157).

References

[1] Huh, J. H.; "An efficient solitary senior citizens care algorithm and application: considering emotional care for big data collection," Processes, MDPI, 2018, 6(12), 1-21

[2] Ng, J., Clay, S. T., Barman, S. A., Fielder, A. R., Moseley, M. J., Parker, K. H., Paterson, C.; Maximum likelihood estimation of vessel parameters from scale space analysis. Image and Vision Computing, Elsevier, 2010, 28(1), 55-63.

[3] Holzinger, A.; "Interactivemachine learning for health informatics: when do we need the human-in-the-loop?," Brain Informatics, Springer, 2016, 3, 119-131

[4] Jordan, M.I.; Mitchell, T.M.; "Machine learning: Trends, perspectives, and prospects," Science, American Association for the Advancement of Science, 2015,263 349, 255-260.

[5] LeCun, Y.; Bengio, Y.; Hinton, G.; "Deep learning," Nature, Nature Publishing Group, 2015, 521, 436-444.

[6] Bayes, T.; "An essay towards solving a problem in the doctrine of chances," Studies in the History of Statistics and Probability 1970, 1, 134-153.

[7] Hastie, T.; Tibshirani, R. & Friedman, J.; "The Elements of Statistical Learning; Data Mining," Inference and Prediction, 2008.

[8] Murphy, K.P.; "Machine learning: a probabilistic perspective," MIT press, 2012.

[9] Man-Kyu Huh, Hong-Wook Huh.; "Genetic diversity and population structure of wild lentil tare," Crop Science, 41.6 (2001): 1940-1946.

[10] M.K Huh., H. W. Huh.; "Allozyme variation and genetic structure of Amorpha fructicosa L population in Korea," Korean Journal of Genetics, 1997, 19.1, 39-47.

[11] M.K Hur., H.W Huh.; "Allozyme variation and population structure of Chimaphila japonica in Korea," Genes & genetic systems, 1998, 73(5), 275-280.

[12] Wu, X.; Fan,W.; Peng, J.; Zhang, K.; Yu, Y.; "Iterative sampling based frequent itemset mining for big data," International Journal of Machine Learning and Cybernetics, 2015, 6, 875-882.

[13] Kim, H. K., So, W. H., Je, S. M.; "A big data framework for network security of small and medium enterprises for future computing," The Journal of Supercomputing, Springer, 2019, 75(6), 3334-3367.

[14] Se-Hoon Jung, Jun-Ho Huh.; "A Novel on Transmission Line Tower Big Data Analysis Model Using Altered K-means and ADQL," Sustainability, MDPI, 2019, 11(13), 1-25.

[15] H. Jung, S. Kim, J.M. Gil, U.M. Kim,; "Processing Continuous Range Queries with Non-spatial Selections," Mobile, Ubiquitous, and Intelligent Computing. Springer, 2014. 31-38.

[16] Kwanho In, Seongkyu Kim, Ung-Mo Kim.; "DSPI: An Efficient Index for Processing Range Queries on Wireless Broadcast Stream," Mobile, Ubiquitous, and Intelligent Computing, Springer, 2014, 39-46.

[17] Larsen, Ross.; "Missing data imputation versus full information maximum likelihood with second-level dependencies." Structural Equation Modeling: A Multidisciplinary Journal, Taylor & Francis, 2011, 18(4), 649-662.

A Tensor Visualization Method for Convolutional Neural Network Modeling Support

Ki Sun Park, Keon Myung Lee*

Dept. of Computer Science, Chungbuk National University, South Korea
gisun1000@naver.com, kmlee@cbnu.ac.kr
*Corresponding Author

Abstract. We present a simple prototype of GUI-based convolution neural network modeling environment for tensor visualization. We propose a visualization method to easily understand the data flow and data shape in the process of designing the convolution neural network. In this method, the shape of the tensor between layers is dynamically drawn on the drawing sheet. The visualized tensor allows developers to easily identify and modify the data flow between layers.

Keywords: Tensor Visualization, CNN, Modeling, Deep Learning

1 Introduction

Deep Learning, one of the machine learning algorithms, has been performing amazingly in various fields such as image recognition, video analysis, and natural language processing. In particular CNN, which is useful in finding patterns such as object recognition, and face recognition, is one of the most powerful deep learning algorithms that can be applied to areas that require computer vision. Most of recent models have won ILSVRC competition are based on CNN.

With this incredible performance, CNN is an algorithm that many developers want to apply first in the field of image recognition. Along with the desire to apply CNN, a variety of libraries and frameworks have been opened [1][2]. Despite the existence of these useful tools, there are many challenges for developers to apply CNN model as it requires deep intuitiveness to the network, along with basic knowledge such as how the convolution layer works and the how data flows between layers. Especially, complex hyper-parameters exist such as input and output size, the number of channels, activation functions, and filter size, which must be considered as essential.

Therefore, this paper proposes a visualization method that makes developers easy to understand of data flow in convolutional neural network. The proposed method provides visualized tensor between layers as diagrams. Developers can directly and dynamically check data flows between layers, which can improve their understanding of the network.

2 Related Works

2.1 Convolution Neural Network

Convolution Neural Network, one of Deep Learning algorithms, was introduced in a paper proposed by LeCun in 1989 [3]. Since then, CNN has generalized by Benhnke in 2003[4], and has been simplified in Simard et al. [5]. Especially since 2012, CNN model has dominated the ILSVRC competition. With this amazing performance, CNN has become one of the most popular techniques in the field of computer vision.

CNN is performing amazingly with technological advances such as follows: i) hardware, ii) datasets and benchmarks, and iii) advanced algorithms. First, NVIDIA and AMD released high-performance GPUs in line with the development and demand of the gaming industry, and the development of parallelism capabilities, such as these clusters, makes it possible that weren't possible before. In particular, Moore's Law is currently in progress, and CNN is also under this rule. Next, with the release of refined datasets such as ImageNet and MNIST, anyone can easily validate their models against these datasets. In online competitions like Kaggle, for instance, you can try out your own and others' models. Finally, the development of advanced activation function and optimizer allows us to build more CNN layers [6].

2.2 GUI Based Modeling Tools

Developments have been stimulated the desire to apply these models, so many libraries and frameworks have been proposed to assist them. TensorFlow [1] and Keras [2] are the most popular tools for convolution neural network modeling. Despite these useful tools, there are still limits for developers to use the deep learning models. Because it is essential to learn the grammar of a particular programming language such as Python.

In our last study [7], we proposed a GUI-based deep learning modeling tool that focused on supporting non-expert developers. Developers are able to design a deep learning model by placing and connecting the deep learning layer to the sheet via drag-and-drop method. As a result, a deep learning model is drawn in the GUI and converted into an executable Python script. Through the proposed methods in the previous paper, we have been able to improve the access and convenience of developers, but how to visualize the data flow has not yet been considered.

3 Visualization of Tensors between Convolution Layers

The expansion of tensor size between each layer can be seen as a decompression process, while the reduction process can be seen as a compression process. In other words, it can be regarded that how the data extracted depends on how the size of the tensor is maintained. Therefore, when constructing a convergence neutral network, attention should be paid to the data flow between each layer, namely the flow of the tensor. In order to focus on the data flow between each layer, the following factors should be considered: i) input size, ii) the number of channels, iii) filter size, iv) paddings, and v) strides. By these factors, the shape of the output tensor is determined.

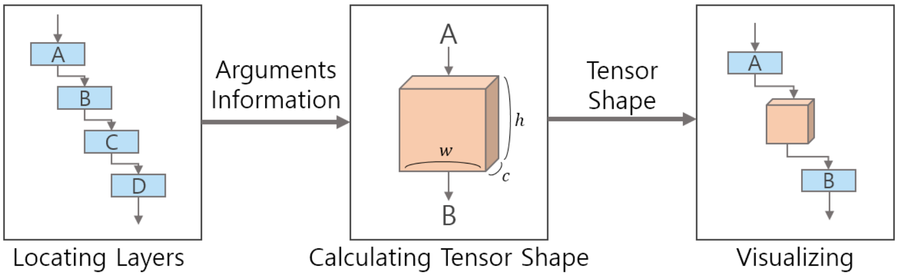


Fig. 1. The Process of Proposed Visualizing Method

Fig.1 shows the process of the proposed visualization method. First, the arguments of convolution layer are set up in the GUI-based modeling tool. The output tensor is calculated by the set values in this process. Based on the calculated output tensor, it is represented dynamically in the modeling tool. Developers can modify the arguments of each layer, after then the shape of tensor is drawn again.

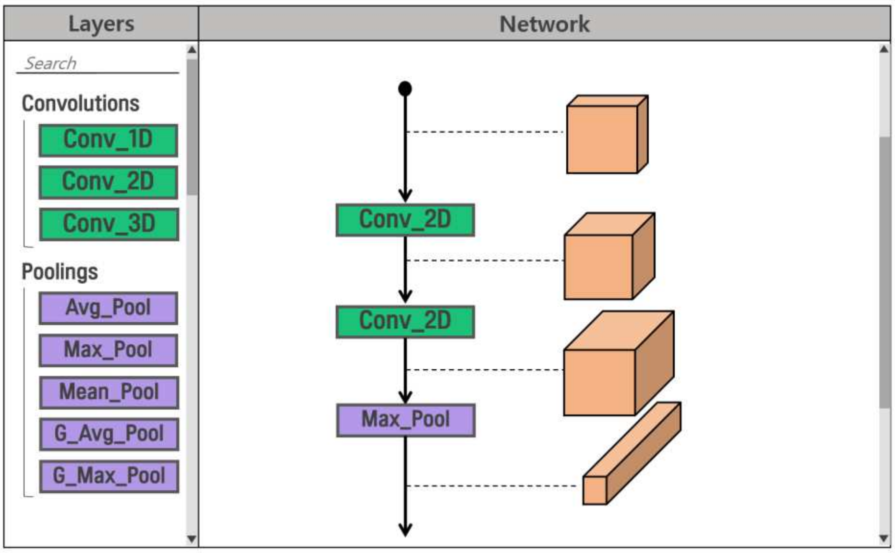


Fig. 2. An Example of Tensor Visualization

In Fig.2, the diagrams in left side indicate the adoptable layers, and it will be located on the right sheet via drag-and-drop manner. On the right side, the connected diagrams indicate the layered architecture. And then the visualized tensor shapes are represented between layers.

4 Conclusions and Future works

In this paper, we proposed a method to visualize the tensor between convolution layers. The proposed method dynamically draws the shape of the tensor by calculating the output size as the internal arguments of the previous layer. This method makes it easy for developers to understand the data flow between layers. In the next research, we plan to study the method in which can visualize the weights and its extracted features between the convolution layers.

Acknowledgments. This research was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (Grant No. NRF-2017M3C4A7069432).

References

1. Martín, A. et al.: TensorFlow: A System for Large-Scale Machine Learning. In: Operating Systems
2. Keras, https://keras.io/
3. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D.: Backpropagation Appled to Hadwritten Zip Code Recognition. In: Neural computation (1989)
4. Benhnke, S.: Hierarchical neural networks for image interpretation. In: Springer, vol. 2766 (2003)
5. Simard, P. I., Steinkraus, D., Platt, J. C.: Best practices for convolutional neural networks applied to visual document analysis. In: Icdar, vol. 3 (2003).
6. Chollet, F.: Deep learning with Python. (2017).
7. Park, K. S., Hwang, K. S., Lee, K. M.: Building Block Identification from Deep Neural Network Codes for Deep Learning Modeling Support. In: International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 3C, pp. 370--375 (2019)

A Novel on Bicycle Big data Prediction Model Using Public Data and Regression Algorithm

Se-Hoon Jung*, Jun-Ho Huh**, Chun-Bo Sim***
*Assistant Professor of **School of Connection Major (Bigdata Convergence), Youngsan University of Yangsan**, Republic of Korea
*Email: shjung@ysu.ac.kr
Assistant Professor of Dept. of Software, **Catholic University of Pusan, Republic of Korea
**Email: 72networks@cup.ac.kr or 72networks@pukyong.ac.kr
*** Professor of School of Multimedia, **Sunchon National University**, Republic of Korea
***Corresponding Author Email: cbsim@sunchon.ac.kr

Abstract. As air pollution becomes a global issue, each city in many countries, public bicycles are installed and operated in major cities of various countries in order to decrease environment pollution and solve traffic problems, additionally operational data of public bicycle stored in real time. But, operation data of bicycle has not been used. Therefore, the purpose of this study is to design a prediction model for forecasting daily rent amount of public bicycle in Suncheon using weather data of KMA (Korea Meteorological Administration) and real time air information of KECO (Korea Environment Corporation). In order to design a prediction model, it is elicited from regression analysis and multiple regression analysis among the amount of daily rent, weather data of the KMA, and real time air pollution degree.

Keywords: Public-Bicycle, Environment Data, Data Analysis, Multiple regression

1. Introduction

As air pollution becomes a global issue, the world's major cities are installing and operating public bicycles in each city to solve not only environmental pollution but also traffic problems. In 2008, public bicycles were first introduced in Changwon under the name "Nubiza", according to the Korea Transport Institute's survey on bicycle use in 2016, it was found that public bicycles were operated in 12 cities and cities nationwide by 2016. Also, the management information data of operated public bicycles is managed by each local government or private company. Public bicycle operation data has an average of about 10,000 pieces of data per day in Suncheon city, and daily average amount of rental is about 400times at around 38 terminals as of April 2017. And in other cities, public bicycle data is stored in real time, but it is not utilized. Accordingly, in this study, we design a prediction model to predict the daily rental amount of public bicycle in Suncheon using the operating information data of Suncheon public bicycle, daily weather data of KMA, and hometown air information data of KECO [1-5].

2. Related Research

2.1. Water Quality Prediction and Water Pollution Source Change

From the viewpoint of the existing water resources management, the integrated management considering various factors such as water quality, aquatic ecosystem, and environment, and a model for predicting the occurrence of algae with interest in the recent steadily increasing algal bloom [6]. To develop the algal prediction model in the proposed study, we use the average of the meteorological factors, hydrological factors, chemical and biological factors, monthly cumulative precipitation, flow, DO, TN, TP. In addition, ARIMA and a multivariate time series model (VECM) were constructed to predict the effect of algae generation on the prediction accuracy of the two models. As a result, Chl-a was found to have the greatest influence on the occurrence of algae. ARIMA and VECM predictions showed that the ARIMA model was relatively high, and both models showed a tendency to reduce the error rate during the winter season compared to the summer season. In order to analyze whether the change of water pollution source of Mihocheon which is one of the Geumgang tributaries affects BOD (Biochemical Oxygen Demand), the QULA-MEV model developed by the National Institute of Environmental Research and the pollution source of 2020 in Chungcheongbuk-do (Population, Livestock, , The state of progress of urbanization) was used [7]. According to the results of the study, the BOD emission of Miho Stream was estimated to decrease to 2,094.6 kg / day in 2020 compared to 2012 due to decrease of population, decrease of cows, cattle, pigs, increase of poultry, It was confirmed that the discharge load decreased as the pollutant sources in the Miho stream watershed decreased.

2.2. Prediction of movie evaluation using user's preference

This study is a system for analyzing user 's preference, watching a specific movie and predicting evaluation of movie. Using the scores and dates of 17,700 movies rated by Netflix customers, analyzed the user's propensity and predicted the ratings after watching a certain movie. Using data mining to classify the prediction results into nine categories and present the prediction results. It is did not consider data that would reduce the accuracy of predictions using all the data collected during the prediction model generation. Therefore, although the predictive power and the significance of the prediction model may be high, there is a problem that overfitting may occur.

3. Novel on Analysis and Prediction Model of Transmission Line Tower Big Data

3.1 Prediction model generation total flow

Fig. 1 is a total flow for generating the prediction model proposed in this study. Firstly, collect public bicycles, real-time air information, and daily weather data.

Secondary, before analyzing the data, preprocess the collected data to improve the accuracy. Thirdly, to generate an appropriate prediction model, the relationship between the amount of daily rental and the real-time air information and daily weather information is confirmed through multiple regression analysis. Fourthly, based on the results of multiple regression analysis, a daily bicycle rental amount prediction model is created.

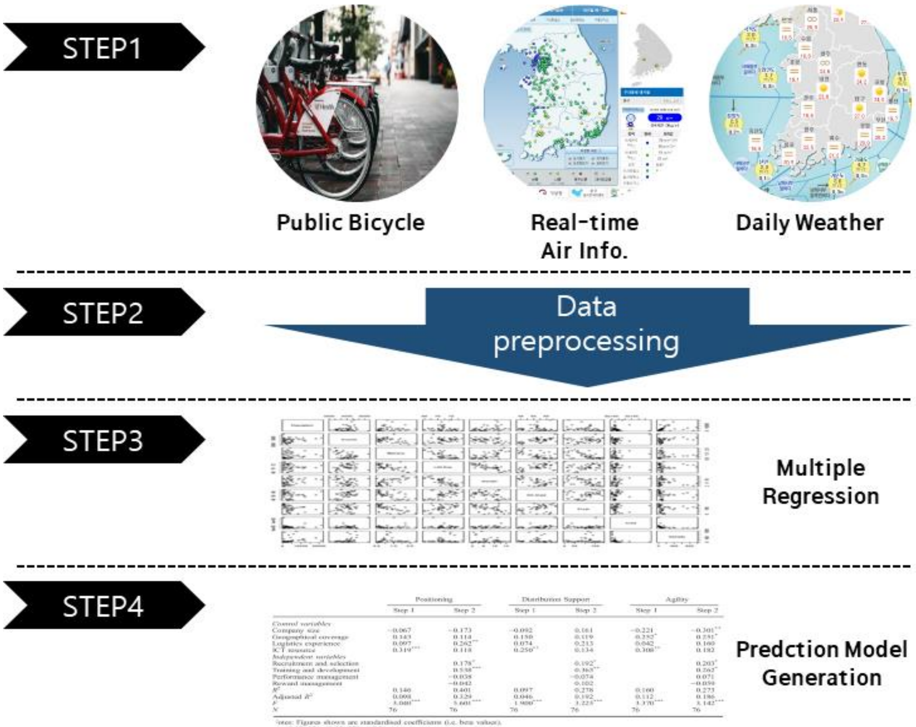


Fig. 1. Structure Diagram of Proposed Model

3.2 Data analysis environment

To derive a prediction model for verifying the above hypothesis, it is necessary to analyze the relationship of data. Table 1 shows the development environment for the design and verification of the proposed prediction model. Currently, R and Python are used as the data analysis language. R is a language optimized for statistical calculations. It has the advantage of providing various data types and packages needed for statistical calculation, but it has a disadvantage of limited use range. Python offers a variety of statistical calculation packages, but it can be used more universally than R. In the proposed study, R is used as a data analysis language and Python is used as a prediction model verification and evaluation language.

Table 1. Development Environment

Classification	Details
CPU	Intel Core i7-6800K 3.40GHz
RAM	24GB
OS	Windows 10 Pro
IDE	R-Studio, PyCharm

Data to be used for data analysis can be broadly divided into public bicycles, KMA, and KECO. Table 2 is the data set used for data analysis. Firstly, the public bicycle data includes public bike operation information including the rental time, the return time, and the event indicating the purpose of the rental. Secondly, the Meteorological Agency data is the past weather information of the relevant region, which includes the daily minimum temperature, maximum temperature, average temperature, daily precipitation, and mean cloud, excluding average cloud. Thirdly, Korea Environment Corporation data includes air pollution information per province, among which PM10 and PM2.5, which means fine dust index, are used.

Table 2. Used data set

Data	Detail Description	Unit
Public bicycle	Rental time	Time(s)
	Return time	Time(s)
	Event	Bool
KMA	Minimum temperature	℃
	Highest temperature	℃
	Average temperature	℃
	Daily precipitation	mm
KECO	PM10	$\mu g/m^3$
	PM2.5	$\mu g/m^3$

3.3 Prediction model generation

To verify the hypothesis of 3.1, we use the Korea Environment Corporation data, the meteorological data, and the operating information data of 'Onnuri bicycle' recorded from December 20, 2015 to December 14, 2017. To derive the analytical model, various models are created for the comparative evaluation after the regression analysis by setting the daily rental amount(RA) of the bicycle as the dependent variable and the daily waiting data of the meteorological agency as the independent variable. Formula 1 is a predictive model derived from multiple regression analysis.

$$RA = 29.2582 - (15.804*TEMP_{avg}) + (26.2790*TEMP_{HIGH}) - (3.288*PT) - (1.330*PM_{10}) - (3.091*PM_{2.5}). \quad (1)$$

4. Result

Figure 2 show multiple regression results. To apply the principal component analysis to the regression analysis, the matrix data is subjected to principal component analysis on the existing data. Multiple regression analysis was used to exclude independent variables that did not meet the significance level of 0.05 to find significant independent variables for dependent variables. As a result, p-values of average temperature, maximum temperature, precipitation, fine dust, and ultrafine dust data were measured to be less than 0.05, which was confirmed to be a significant variable to the amount of bicycle lending. Figure 3 shows the distribution of data, normal distribution, detection of anomalous points, and degree of deviation of data.

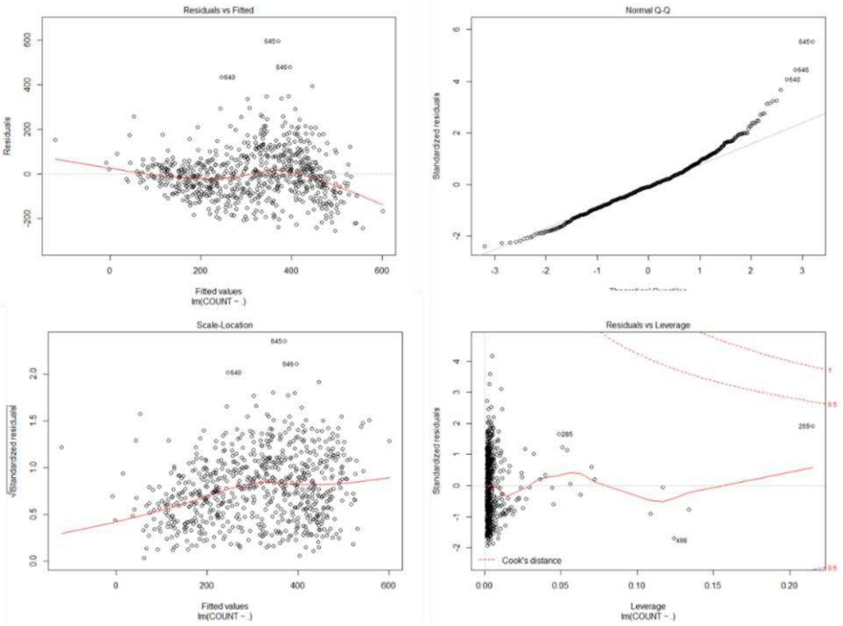


Fig. 2. Multiple regression analysis diagnostic graph

5. Conclusion

In this study, we propose a forecasting model of daily rental amount of public bicycle using weather data of Korea Meteorological Agency and Korea National Housing Corporation's waiting information data. In the future, we will generate various prediction models using the results of multiple regression analysis and estimate the accuracy by deriving the optimal model through comparative evaluation. In this study, bicycle rental amount per day was predicted by using cross validation, principal component analysis and multiple regression analysis. In the future, we will conduct research on forecasting the amount of rental per hour by utilizing public bicycle data of various cities as well as public bicycle data of two years.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT)(NRF-2019R1G1A1002205).

References

[1] Korea Transport Institute's survey, https://www.ktdb.go.kr
[2] Suncheon city public bicycle rental amount, http://bike.suncheon.go.kr
[3] Seoul Metropolitan Facilities Management's Comprehensive Status of Public Bicycles, https://www.sisul.or.kr
[4] Climate data of Korea Meteorological Agency, http://www.weather.go.kr
[5] Hometown air information data of KECO, https://www.keco.or.kr/kr
[6] J. H. Song, "Prediction and Characterization Analysis of River Water Quality using Multivariate Time Series Models", Master's thesis, Graduate School, University of Seoul, (2017). (In Korean)
[7] T. K. Kim, "Water Quality Prediction of Miho Stream according to Pollution Source Change", Journal of Industrial Science Researches, vol. 35, no. 2, (2018), pp. 1-10.
[8] H. Jung, S. Kim, J.M. Gil, U.M. Kim,; "Processing Continuous Range Queries with Non-spatial Selections," Mobile, Ubiquitous, and Intelligent Computing. Springer, 2014. 31-38.
[9] Kwanho In, Seongkyu Kim, Ung-Mo Kim.; "DSPI: An Efficient Index for Processing Range Queries on Wireless Broadcast Stream," Mobile, Ubiquitous, and Intelligent Computing, Springer, 2014, 39-46.
[10] Man-Kyu Huh, Hong-Wook Huh.; "Genetic diversity and population structure of wild lentil tare," Crop Science, 41.6 (2001): 1940-1946.
[11] Se-Hoon Jung, Jun-Ho Huh.; "A Novel on Transmission Line Tower Big Data Analysis Model Using Altered K-means and ADQL," Sustainability, MDPI, 2019, 11(13), 1-25.
[12] Kim, H. K., So, W. H., Je, S. M.; "A big data framework for network security of small and medium enterprises for future computing," The Journal of Supercomputing, Springer, 2019, 75(6), 3334-3367.

Create List of Stopwords and Typing Error by TF-IDF Weight Value

Woo-seok Choi[1], Ki-cheol Yoo[2], Sang-Hyun Choi[2],

[1] Department of Bigdata, Chungbuk National University,
Cheongju, South Korea
[2] Department of Manangemnt Information System, Chungbuk National University,
Cheongju, South Korea
{cdt3017}@naver.com, {ryugami07, chois}@cbnu.ac.kr

Abstract. On these days, development of SNS generate huge text data. It is most important things to remove the meaningless words, stopwords and typing error to analyse text data. In English, it grew rapidly to create stopwords dictionary. However, there are few researchs in Korea for Korean language. In this research, we suggest way to firter stopwords and typing errors out by words importance with TF-IDF algorithm. First, calculate TF-IDF value from collected data. Second, decide criteria to separate to two groups by TF-IDF value and transform to n× 2 matrix. Third, calculate accumulative frequency of TF-IDF weight. In this way, new accumulative frequency is gotten without stopwords and typing error. Furthermore, this method can be used in both language : Korean and English. without creating stopwords dictionary.

Keywords: TF-IDF, Text Mining, Stopwords, Preprocessing

1 Introduction

1.1 Purpose and Background of Research

In this research, we suggest efficient way to filter out meaningless word such as stopword, typing error, article, preposition, postposition, conjunction and just often used word by its data type in text mining.

In text mining, researchers don`t analyse normal structured data, but analyse unstructured data in many kinds of format. That`s why researcher must process the data to enable to analyse. Especially, stopword and typing error must be filter out in preprocessing to avoid affect to entire output. For this kind of pre preprocessing, two ways are often used. Make stopword dictionary or process by each word frequency [1].

However, it is different between Korean and English. There are few researches for Korean stopword dictionary. It is really hard to find Korean stopword list [2] So, in this research, we suggest way to filter out the stopwords and typing error by its frequency and weight.

2 Related Research and Method

2.1 Existing Research

Gill-Hohyun(2018) extracted 10,000 of morpheme and removed important role words in sentence, formal morpheme with high frequency and meaninglessness. And Gill suggested draft proposal of Korean stopwords dictionary made by the words: high frequency and meaninglessness [3].

Lee-Minsik, Lee-Hongju (2016) devised Text-Word matrix to make division for customer review; is it worth it or not, to remove the wordes by scarcity and neutrality. And they classified the review in two group: the worth review group, and meaningless review group [4].

2.2 TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency) is kind of weight model for information searching and text mining. It is statistical value which is judged each word is how important in each text file. It is used for extracting key words from documents, deciding web research key words ranking, and comparing similarity between documents.

TF (term frequency) is value of how often appear each words, DF(document frequency) is the value about the particular words is in the documents or not. IDF is inverse of DF. So, TF-IDF could be expressed by multiply of TF and IDF.

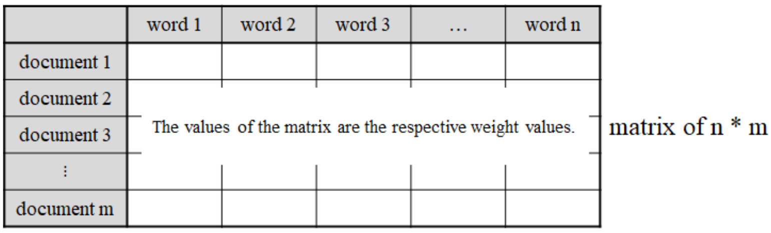
Table 1. TF-UDF Weight Model.

TF (term frequency)	$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: *the count of words 'ti' in document 'dj'* $\sum_k n_{k,j}$: *the count of every appearance of every word in the documents 'dj'*
DF (document frequency)	$idf_i = \log \frac{\lvert D \rvert}{\lvert d_j \lvert t_{i \in} d_j \rvert}$ $\lvert D \rvert$: the number of documents belonged in documents set $\lvert d_j \lvert t_j \in d_j \rvert$: the number of documents which have the words 't_j '
TF-IDF	$TFIDF_{i,j} = tf_{i,j} \times idf_i$

3 Related Research and Method: Accumulative Frequency with TF-IDF Weight

3.1 Problem of TF-IDF

TF-IDF is multiplied value of TF and IDF. And it is expressed by the matrix: the number of 'unique' words * the number of documents file. For example, the number of words which is removed overlap is 4 and the number of documents (sentence) is 10, then as a result, the matrix has 4 * 10 size.



	word 1	word 2	word 3	…	word n
document 1					
document 2	The values of the matrix are the respective weight values.				
document 3					
⋮					
document m					

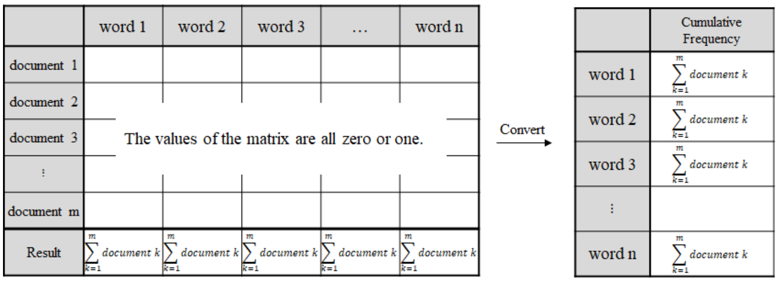
matrix of n * m

Fig. 1. Size of TF-IDF Result.

Such a small data like example, there are no problem. But the more words, the more data storage. The data storage needs are growth exponentially. For example, the example has 0.1millions of sentences and 1millions of words, it becomes 100,000 * 1,000,000 matrix and it means the number of 100,000 * 1,000,000 of TF-IDF weight.

3.2 Transformation of TF-IDF Result

In this research, the Hyper parameter(H) was decided to solve the exponential growth of TF-IDF value. In the documents, the number of the words more than particular number is 0, or it couldn`t, then the words get 1. It means, put every weight which size is m* n in the model and transform the weight between 0 or 1. After that, sum every result of the documents and make it as matrix which size is n* 2.



	word 1	word 2	word 3	…	word n
document 1					
document 2					
document 3	The values of the matrix are all zero or one.				
⋮					
document m					
Result	$\sum_{k=1}^{m} document\ k$	$\sum_{k=1}^{m} document\ k$	$\sum_{k=1}^{m} document\ k$	$\sum_{k=1}^{m} document\ k$	$\sum_{k=1}^{m} document\ k$

	Cumulative Frequency
word 1	$\sum_{k=1}^{m} document\ k$
word 2	$\sum_{k=1}^{m} document\ k$
word 3	$\sum_{k=1}^{m} document\ k$
⋮	
word n	$\sum_{k=1}^{m} document\ k$

Fig. 2. Transform of TF-IDF and summate and Transform to n * 2.

In other words, TF-IDF weight value means importance of each words in each document. In this research, we decide criteria for hyperparameter H made by TF-IDF weight and classified 'meaningful' or 'meaningless' by H value. After that, summate every 'meaningful' word and calculated new accumulative frequency by TF-IDF weight.

3.2 Accumulative Frequency with Weight: Remove Stopwords

TF-IDF method gives low weight to meaningless words. Automatically, stopwords got low weight. It means, most of stopwords are excepted from accumulative frequency. It is possible to check the ranks without stopwords, typing error, meaningless words in this way.

4 CONCLUSION

5.1 Conclusion of Research

As we said, accumulative frequency by TF-IDF weight method is the way to classify meaningless words by hyper parameter(H) from TF-IDF weight. In existing research, researcher should create the stopwords dictionary first to remove the meaningless words. But this research has huge advantage that it is possible to remove the stopwords without other processing in very efficient way.

In this research, even we only used our method to only nouns, this method also can be used for verbs and adjectives. Especially, when doing sensitivity analysis by using verbs and adjectives, this method is good pre-process method to remove meaningless words.

5 REFERENCE

1. Miah K, Min S.: A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis. Korea Intelligent Information System Society. 18, 53--77 (2012)
2. Hohyun K.: The Study of Korean Stopwords list for Text mining. urimalgeulhakhoe. 78, 1--25 (2018)
3. Bongjun C. Hangjoo L.: A Generation and Matching Method of Normal-Transient Dictionary for Realtime Topic Detection. The Journal of KINGComputing. 13, 7--18 (2017)
4. Minsik L. Hongjoo L.: Increasing Accuracy of Classifying Useful Reviews by Removing Neutral Terms. Korea Intelligent Information System Society. 22, 129--142 (2016)

Analysis of zone-based registration with three zones

Hee-Seon Jang[1], Jihee Jung[2], Jang Hyun Baek[3*]

[1]Department of Convergence Software, Pyeongtaek University, Pyeongtaek, 17869, Korea
[2]CAMTIC Advanced Mechatronics Technology Institute, Jeonju, 54852, Korea
[3*]Department of Industrial & Information Systems Engineering
Chonbuk National University, Jeonju, 54896, Korea, jbaek@jbnu.ac.kr

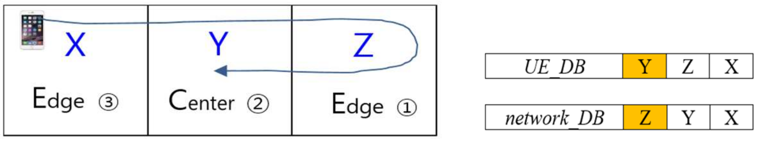
Abstract. The location of a user equipment (UE) should always be maintained to connect an incoming call in mobile cellular network. Many location registration schemes have been proposed but most mobile cellular networks have adopted the zone-based registration because of its good performance. Even though the recommendation says that a UE stores several zones in the zone-based registration, a UE keeps a single zone in most of the mobile cellular networks. Some studies have been performed on the zone-based registration with several zones but most of these studies consider the zone-based registration with only two zones. In this study, by using the embedded Markov chain model, we give a simple mathematical model for the zone-based registration with three zones. In addition, by using the research results on the zone-based registration with one, two and three zones, we suggest the optimal management scheme of the zone-based registration. Since most mobile cellular networks adopt the zone-based registration, these research results can be used for enhancing the performance of current real mobile cellular networks.

Keywords: location registration, paging, zone-based registration, embedded Markov chain.

The location of a user equipment (UE) should always be kept in order to connect an incoming call to the UE in a mobile cellular network. Several methods of registering locations have been proposed. Most mobile cellular networks have adopted the zone-based registration due to its good performance and ease of implementation. In general, the mobile cellular network is composed of many zones that is made up of several cells. A UE should register a new location information to the network database (DB) every time it enters a new zone, which is called *location registration*. When an incoming call arrives to the UE, the network pages the UE to all of the cells in the UE's zone to find the cell to which the UE belongs and connect the call. This process is called *paging*. Generally, as the number of cells in a zone decreases, the registration load increases but paging load decreases. Therefore, it is necessary to determine the optimal operating scheme of a zone-based registration by considering the trade-off between the registration and paging loads.

While the recommendation for mobile cellular networks says that a UE can store several zones in zone-based registration, the UE stores a single zone in the actual mobile cellular network. Therefore, some studies have been performed on the zone-based registration with several zones but most of the studies consider the zone-based registration with two zones. In this study, by using the embedded Markov chain model, we present a simple but accurate mathematical model for the zone-based registration with three zones (3ZR).

Many studies on location registration employ a symmetric random-walk mobility model. In this case, it is assumed that a UE moves to all neighboring zones with the same probability. However, it is easy to see that this assumption is not realistic if we consider the mobility of the actual UE. In general, the next entering zone of the UE is tightly related to the current zone, which means that the assumption of UE's symmetric random walk mobility is not correct. In our study, to solve this problem, the probability of returning to the previous zone is assumed to be θ to reflect the dependency between the current zones and next entering zone. In a square zone environment as shown in Fig. 1, θ would normally have a value greater than 0.25. Note that θ=0.25 means a symmetric random walk model.



Network can't know UE returns to Y since there is no registration

Fig. 1. Edge and center zones in 3ZR

To explain the embedded Markov chain model, let us consider the relation of the three zones for 3ZR shown in Fig. 1. Among the three zones, let's call the two zones at each end *edge* zones and the middle zone as *center* zone, and mark as E and C respectively. Sometimes, it is useful to mark two edge zones as ① and ③, and the center zone as ②. The edge zone, marked as ①, indicates that the location registration was made there more recently than the other edge zone (marked as ③).

Next, let us define the states E_{ij}, C_{ij} (i=1, 2, 3, j=1, 2, 3) for Markov chain model in the regular case. E and C indicate that the most recently registered zones are the edge and center zones, respectively. The first subscript i (= 1, 2, 3) represents the UE's immediate previous zone, and second subscript j (= 1, 2, 3) indicates the UE's current zone. Now, we can obtain state transition diagram by mapping every event such as cell entrance, call arrival and so on into transition between related states.

Next, using the state transition diagram, we can obtain steady state probability of each state and finally calculate the paging cost and the registration cost for 3ZR. The numerical results for the various situations show that, when the call-to-mobility (CMR) is small, the 3ZR is better than the zone-based registration with one or two zones. On the other hand, when the CMR is very large, the 3ZR is not so good. Finally, the research results on the zone-based registration with one, two and three zones are used to suggest the optimal management scheme of the zone-based registration. Since most mobile cellular networks adopt the zone-based registration, these research results can be used for enhancing the performance of current real mobile cellular networks.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B01014615) and by the Ministry of Science, ICT and Future Planning (2017R1E1A1A03070134).

Instance Segmentation Network Using Dilated Convolution in Parallel

Sung-su Jang[1], Young-guk Ha[1,*]

[1] Department of Computer Science & Engineering
Konkuk University
Neungdong-ro, Gwangjin-gu, Seoul 143-701, Korea
pik1100@naver.com, ygha@konkuk.ac.kr

Abstract. DNN has recently shown good performance in high-level vision tasks such as image classification and object detection. This paper utilizes Faster R-CNN in order to detect image objects efficiently and generate high quality segmentation masks for each instance. The mask utilizes Dilated Convolution. In addition, the accuracy is improved by using a technique that combines various rates of Dilated Convolution in parallel.

Keywords: Deep Learning. Instance Segmentation, Dilated Convolution

1 Introduction

There are three types of segmentation: semantic segmentation, instance segmentation, and panoptic segmentation. Segmentation is one of the core areas of computer vision. It's a high-level problem that not only looks at the photos and classifies objects, but also requires a complete understanding of the scene. Research is being actively conducted in various fields such as autonomous driving and medical fields. In this paper, we will apply several techniques to improve the performance of instance segmentation.

2 Related Works

Fully Convolutional Network (FCN) [1]is a semantic segmentation technology based on CNN. FCN is a model that transforms the Fully Connected layer into 1x1 convolution behind the basic CNN model. Dilated Convolution [3] is a method which forcibly increases the receptive field by adding zero padding inside the filter. A typical CNN is a sequence of conv. and pooling process. However, pooling result in the loss of existing information. To solve this problem, it is appropriate to use Dilated Convolution. Dilated Convolution can use large sized receptive field without polling. It is appropriate to use Dilated Convolution to maintain spatial characteristics.

[*] Corresponding author

3 System Architecture

In Figure 1, the object detected image is entered the input image through Faster R-CNN. [2] Segmentation is then performed in the bounding box. The process proceeds to Dilated Convolution in the bounding box, where the rates are varied and merged in parallel. It is expected that a good result will be obtained if the correction is made with CRF later.

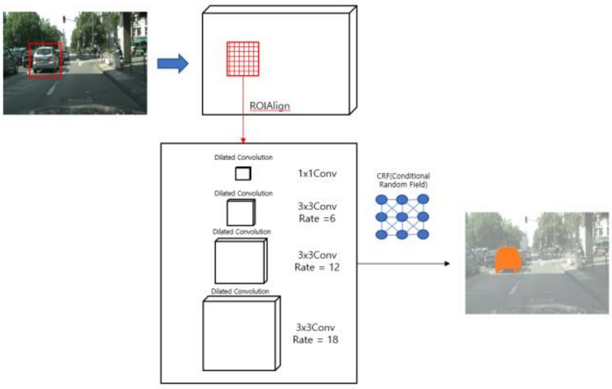


Fig. 1. System Architecture

4 Conclusion

We studied techniques which improve the performance of instance segmentation. And we expect that using dilated convolution in parallel could improve segmenting performance. Future research will verify the performance through various experiments and adjust the network structure to achieve the desired performance.

Acknowledgments. This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program) (20000946 , Development of artificial intelligent computing platform technology for service robots capable of real-time processing of large-capacity, high-performance sensor fusion processing and deep learning) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

References

[1] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[2] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[3] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).

Analysis and Design of Web Interface in Mobile Web Browser for the Low Vision People

Joo Hyun Park[1], Soon-Bum Lim[2]

[1] Dept. of IT Engineering, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 04310, Korea
[2] Research Institute of ICT Convergence, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 04310, Korea
park.joohyun5@gmail.com, sblim@sm.ac.kr

Abstract. Low vision people are familiar with acquiring information through a PC-based web browser using commercially available technology. Recently, as mobile devices become a necessity for the visually impaired, information retrieval and sharing activities are shifting to mobile device environment. However, when the existing PC-based web browser service is applied to the mobile environment, problems such as GUI and TTS control as well as enlargement and selection of context arise. Therefore, this study suggests the development of a mobile web browser that provides context focusing function and user selective TTS service for low vision and elderly people.

Keywords: Low Vision, Web Accessibility, Multimodal Web Interface

1 Introduction

The use of computers and the Internet is essential for informatization, but the disabled and the elderly are having difficulty using the web browser [1, 2]. Low vision people often use Web browsers based on the PC environment using technologies that are already commercially available. However, when applying existing PC-based services to the mobile environment, there are limitations in GUI and TTS problems as well as the expansion and selection of contexts [3]. In addition, the elderly population is steadily increasing, and 86% of visually impaired persons are low vision. It is necessary to develop a mobile browser service that can be used by both the elderly and the low vision. Therefore, we propose the development of a mobile web browser that provides user selected TTS function and context focusing services for low vision and the elderly.

2 Design of Interface

The main functions of the service are classified into structural personalization of menu / function, selective focusing of context, user selective voice service, and color reversal.

First, the structured personalization function of the menu / function displays the main functions of the web browser as icons on the layer created separately from the browser screen. This includes navigation menus that include functions such as home, previous, and next, and personalization menus such as favorites, selection, zoom, and settings. To use this feature, access the tablet by selecting the icon of circle with yellow on black border located at the bottom right of the web browser launch screen.

The contextual selective focus function is to enlarge or play TTS by selecting the menu and content paragraph of the web page after setting the 'select' mode. At this time, the area of the selected context is highlighted so that the user can recognize the selected area. Once you have selected the paragraph you want, press the circle icon at the bottom of the screen and the 'Selection Mode Tool' appears, including the zoom setting, voice setting, backward, confirmation, and exit. These functions allow the user to configure the voice service function as needed. It can also determine whether to use the context widening service, adjust the context enlargement ratio, color inversion, and adjust the repetition and speed of voice services. In addition, the 'back' button is provided so that the mode can be reset at any time during the setting. When the setting is completed, the setting value can be changed at any time by pressing the circle icon.

Finally, the color reversal function reconstructs the color of the interface according to the type of user, which is divided into low vision and presbyopia. In the case of presbyopia, when the context is selected, areas are displayed in red. In the case of low vision, a color reversal screen was added, and the area selected in yellow was displayed.

This service uses Android Studio as a development tool and developed in java, JavaScript, HTML5, CSS language. Also, TTS (Text-To-Speech) API is used for voice service.

3 Conclusion and Future Work

This service is designed to help low vision and elderly people to read the context of web browser and to use basic functions without any difficulties.

In the future, we expect to be able to enhance the usability of daily life activities by providing expansion and voice services in a variety of functions such as shopping, posting, and e-mail services.

Acknowledgments. This research was supported by the National Research Foundation of Korea (NRF-2018R1A4A1025559).

References

1. Park, J.H., Lee, J.W., Lim, S.B., et al.: Design of Voice Annotation System to enhance the quality of life for Visually-impaired Readers. The Asian International Journal of Life Science, Supplement 11, pp 109-121 (2015)
2. Park, J.H., Kim, H.Y., Lim, S.B.: Development of an electronic book accessibility standard for physically challenged individuals and deduction of a production guideline. Computer Standards & Interfaces, Volume 64, pp 78-84 (2019)
3. Park, J.H., Shin, J.E., Lim, S.B., et al.: Mobile Web Browser Multimodal Interface for the Visually Impaired. In: 2019 Korea Multimedia Society Spring Conference, pp. 181—184 (2019)

Centralized Server-based Light Field Display System Using a Head Tracking Camera

Md. Shahinur Alam, Ki-Chul Kwon, Munkh-Uchral Erdenebat, Yan-ling Piao, and Nam Kim*

School of Information and Communication Engineering,
Chungbuk National Univerity,
Chungbuk 28644, South Korea.
*namkim@chungbuk.ac.kr

Abstract. Light field (LF) display is one of the methods to deliver three-dimensional (3D) images by generating the distribution of the light in the space. However, there are some limitations such as a low resolution and a narrow viewing angle. Also, it needs huge computation to display the perfect LF image. In this paper, we propose a method to implement an LF display using a head-tracking camera from a centralized server which can provide a 3D image with higher resolution in less computation cost. The 3D object is captured by a camera array and stored in a central server, then this object is displayed to the end-user according to the viewing direction, tracked by the head tracking camera.

Keywords: Light field display, 3D display, light field image, Multiview image acquisition.

1 Introduction

Light field display technology has become popular due to the advent of 3D Display and capturing devices. There are different kinds of light field display evolved in the last few decades [1]. Recently, an eye-tracking horizontal parallax based light field display has been proposed [2]. But only the horizontal view isn't good enough to feel the real 3D view; that's why we have implemented a head tracking based light field display system which provides both horizontal and vertical motion parallax. However, there are some limitations such as a low resolution and a narrow viewing angle. In order to enhance the above parameters, several researchers have proposed a temporal multiplexed LF display. In that case, a display device with a higher refresh rate is required to prevent the flicker problem. In this paper, we reduced the computation time by using a central server where most of the calculations are done.

2 Methodology

The proposed head tracking based light field display system is shown in Fig. 1. The whole process can be divided into two main parts- LF image acquisition and display system; between that, a centralized server is used to store and deliver the images. Generally, LF display required an enormous number of images to perceive the perfect motion parallax. Therefore, it requires a huge computation in the display end, hence we proposed a centralized server system that reduces abundant calculation in the user end.

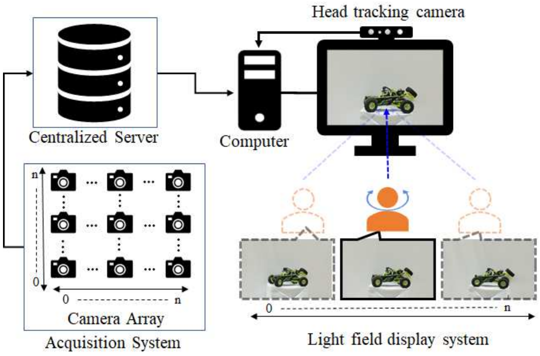


Fig. 1. Head Tracking Based Light Field Display.

The acquisition system shown in Fig.1 has multiple cameras. Here, 0 and n is the leftmost, and rightmost view position, respectively. In our system, there are 71 horizontal and 3 vertical views. The viewing zone and motion parallax are mostly dependent on the number of views. Higher views generate smooth motion parallax. The captured images are joined one by one and stored in the central server as a stereo image.

In the user end, the stereo view images are displayed in a 3D stereo flat panel according to the viewing direction i.e. head position. The head position is tracked by an Intel RealSense SR300 camera. The field of view (FOV) for both acquisition and display system is 30°. It is observed that the combining consequence images interval of four (which is almost 2° view difference) generates a perfect 3D view.

3 Conclusion

In this work, we proposed a light field display method using head tracking technology, and the calculation time reduced by using a centralized server. Therefore, it enhances motion parallax. In this system, the view is limited by 71×3; in the future, we will try to enhance the vertical views by adding more camera.

Acknowledgments. This work was supported by Korea government grant funded by the Korea government (No. GK17C0200, Development of full-3D mobile display terminal and its contents) and the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-2015-0-00448) supervised by the IITP(Institute for Information & communications Technology Promotion).

References

1. J. Geng: Three-dimensional display technologies. In Adv. Opt. Photonics, Vol. 5, no. 4, p. 456, Dec. 2013.
2. L. Yang, X. Sang, X. Yu, B. Liu, B. Yan, K. Wang, and C. Yu: A crosstalk-suppressed dense multi-view light-field display based on real-time light-field pickup and reconstruction. In Adv. Opt. Express, Vol. 26, no. 26, p. 34412, Dec. 2018.

Performance Analysis of two TAL-based Registration in LTE Network

Jae Joon Suh[1], Hee-Seon Jang[2], Gwang Jin Han[3], Jang Hyun Baek[3*]

[1]Department of Industrial & Management Engineering, Hanbat National University, Daejeon, 34158, Korea
[2]Department of Convergence Software, Pyeongtaek University, Pyeongtaek, 17869, Korea
[3]Department of Industrial & Information Systems Engineering
Chonbuk National University, Jeonju, 54896, Korea
*corresponding author: jbaek@jbnu.ac.kr

Abstract. In this study, two tracking-area list (2TAL) based registration is proposed to reduce the signaling cost for registrations in current TAL-based registration in mobile cellular networks. In the proposed scheme, the implicit registration and 2-step selective paging is adopted. Computer simulations are performed to analyze its performance and to compare with original TAL-based registration. Through various numerical results, it is observed that 2TAL-based registration yields better performance as the call-to-mobility ratio has a higher value.

Keywords: location registration, paging, TAL-based registration, 2TAL-based registration, mobility management.

The current location of a user equipment (UE) should be constantly tracked by the mobile cellular networks to connect the incoming calls to the UE. The location management includes two basic operations, location registration (LR) and paging. LR is the process that enables the UE to register its location in network databases, whenever it enters a new location area. Paging is the process through which, when a call arrives, the network finds the cell to which the UE belongs to connect the incoming call to the UE.

Many studies have been performed to reduce the signaling cost for the location management. Deng *et al*. proposed the tracking-area-list (TAL) based registration in which the UE registers its new location when it enters a new TAL that is composed of several tracking-areas (TA, a group of cells). In current TAL-based registration, whenever a UE enters a new TAL, it registers its location so that the TA of entered cell becomes the center TA of new TAL, which is called central policy.

This central policy can reduce the possibility that the UE quickly exits the new TAL. However, under TAL-based registration with central policy, a UE registers its location more frequently than other registration schemes. To reduce the frequent registrations, in this study, we propose 2TAL-based registration, in which a UE can have two TALs. A UE and network store old-TAL and new-TAL simultaneously. In this case, when the UE goes back to the previous TAL stored in the UE, it does not need to register. On the other hand, since a UE does not register when it goes back to the previous TAL, the network may not know its correct location, and the paging cost can increase. In our 2TAL-based registration, the implicit registration by outgoing calls and 2-step selective paging are adopted. Through the implicit registration, the UE's location can be updated without additional LR costs. And, in 2-step paging, the cells in the new-TAL are paged first, and the rest of the cells in the old-TAL are paged next if there is no response from the UE. To analyze the registration and paging cost, we perform computer simulations using RAPTOR according to the flowchart as shown in Fig. 1.

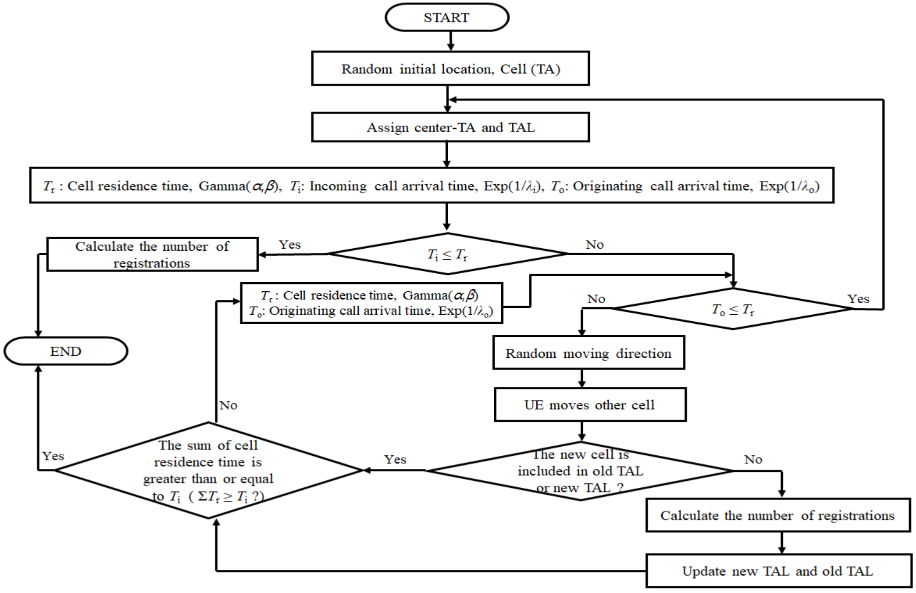

START
Random initial location, Cell (TA)
Assign center-TA and TAL
T_r : Cell residence time, Gamma(α,β), T_i: Incoming call arrival time, Exp($1/\lambda_i$), T_o: Originating call arrival time, Exp($1/\lambda_o$)
Calculate the number of registrations
Yes
$T_i \leq T_r$
No
T_r : Cell residence time, Gamma(α,β)
T_o: Originating call arrival time, Exp($1/\lambda_o$)
No
$T_o \leq T_r$
Yes
END
Random moving direction
UE moves other cell
No
Yes
The sum of cell residence time is greater than or equal to T_i ($\Sigma T_r \geq T_i$?)
Yes
The new cell is included in old TAL or new TAL ?
No
Calculate the number of registrations
Update new TAL and old TAL

Fig. 1. Simulation model of 2TAL-based registration

The simulation consists of the following three main functional modules, assuming the square-shaped cell and TA.

- In the *initialization* module, a cell (in a TA) is randomly selected, and the center-TA (in new-TAL) is assigned. It is assumed that the cell residence time (T_r) follows gamma distribution and interarrival time of incoming calls (T_i) and originating calls (T_o) follow exponential distributions.
- In the *moving* module, the UE enters new cells with random direction after staying during T_r in the current cell.
- The *TAL list update* module inspects whether the new cell ID exists in the memory (in old-TAL or new-TAL). If the cell doesn't exist in the memory, the update process is performed (new-TAL is newly organized, and old-TAL is updated). The number of registrations is computed in this module.

Through computer simulations under various situations, the total signaling cost of 2TAL-based registration is evaluated and compared with original TAL-based registration. It is expected that 2TAL-based registration is superior to original TAL-based registration as UEs' call-to-mobility increases assuming the implicit registration and 2-step selective paging. The results will be appeared in the presentation in details.

Improved Randomized Input Sampling for Explanation of Black-box Models using Segmentation

Ho-rim Park[1], Kyu-hong Hwang[1], Young-guk Ha[1,*],

[1] Department of Computer Science & Engineering,
Konkuk University.
120 Neungdong-ro, Gwangjin-gu, Seoul, Korea
ghfla543@gmail.com, gfvxgd2k@naver.com, ygha@konkuk.ac.kr

Abstract. With the development and practical use of deep neural network, it has a great influence on data analysis and decision making in real life. However, the decision making process is still difficult to explain to the end user and is not clear. In this paper, we propose a system architecture that shows the importance map of the location to be concentrated through the segmentation of super pixel and random pixel of RISE.

Keywords: Deep Learning, Super pixel, Segmentation, Explanation of Deep Learning Model

1 Introduction

Recent advances and successes in deep neural networks have led to significant growth in AI. However, it is generally not clear how the network makes a decision, how certain it is, and whether it can be trusted. In particular, in areas such as autonomous driving, where network decision making can have serious consequences, clarity of decision modeling is important. This paper deals with the problem of providing an explanation for the determination of Explainable AI, that is, an AI model.

2 Related Work

Superpixel[1] is one of many techniques used in image preprocessing. It combines pixels with low level information such as color to make large pixels. Therefore, the effect of greatly reducing the elements constituting the image can be obtained.

The conventional method for estimating pixel dominance of RISE[2] is to search the base model by sub-sampling the input image through the results of random mask and masking image without accessing parameters, features, and gradients. The final importance map is generated from a combination of probability and random mask predicted by the base model from the masking image. This approach estimates the main effect without accessing the internal structure of the base model and can be used for any base model.

[*] Corresponding author

3 System Architecture

Unlike conventional RISE, before applying a random mask to an image, after the segment image is generated through the super pixels, the intensity of the random segment pixels is reduced to zero. The random segment mask image becomes the input image, the mask is randomly masked pixel by pixel, and the output probability is derived by the black box F. The importance map is generated through the combination of the output probability and the random binary linear. Figure 1 shows the system architecture showing the process.

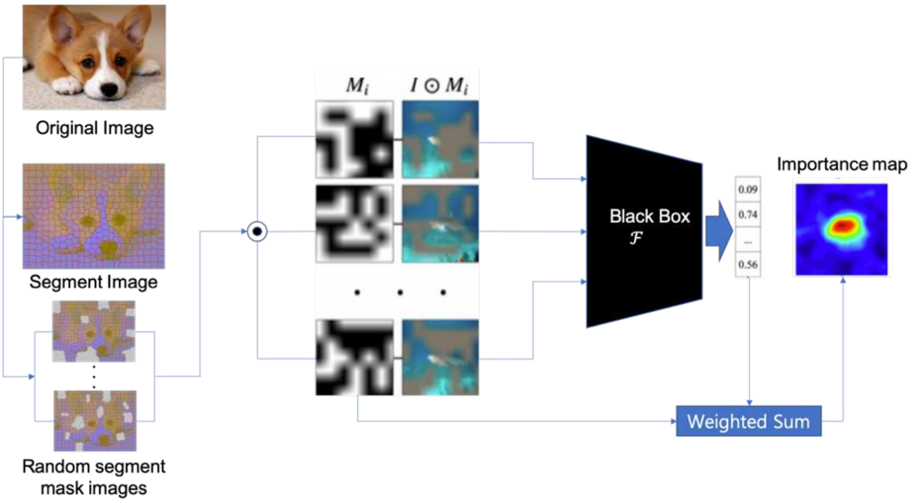


Fig. 1. System Architecture

4 Conclusion

By working with random segmentation, we can derive a clearer importance map, suggesting that this is a clearer and explanation of the black box. Future work is to develop and experiment with the system structure.

Acknowledgments. This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MEST)(No. 2018R1A2B2006154)

References

1. Radhakrishna Achanta, Appu Shaji, Kevin Smith, Pascal Fua, and Sabine Susstrunk, “SLIC Superpixels”, EPFL 2010
2. Vitali Petsiuk, Abir Das and Kate Saenko, “RISE : Randomized Input Sampling for Explanation of Black-box Models”, British Machine Vision Conference 2018

Remote controlled of Door-lock System using Arduino Board

Seol So[1], Ja-Yeon Jeon[2], Soon-Bum Lim[3]

[1,2] Dept. of IT Engineering, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 04310, Korea
[3] Research Institute of ICT Convergence, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 04310, Korea
jyjeon@gmail.com, sblim@sm.ac.kr

Abstract. In recent years, housing intrusion and theft have been raised as major problems. In this project, we implemented a system that can use WIFI communication to confirm the front door from outside, strengthen security, check visitors and open and close doors. The system consists of two Arduino boards that support WIFI communication, CCTV, alarm sound generator and lock. We hope that through this project we can reduce the risk of crime such as burglary, theft and robbery.

Keywords: Remote control, Arduino, Door-lock

1 Introduction

Recently, the family structure has shifted from the large family structure to the family structure of two or less people, and the number of vacant houses in the daytime has increased. Therefore, it has been pointed out that housing intrusion and theft are a major problem. This project has designed and implemented a system for remotely opening a door lock using a smart phone application to prevent password exposure when a customer visits in absence.

2 Design of System

This system is provided with two Arduino boards supporting WIFI communication, a CCTV, an alarm sound generator and a lock device. The door lock function was implemented in C++ language, and the smartphone application was implemented in Java Script language. The system model of this project appears to be [Figure 1].

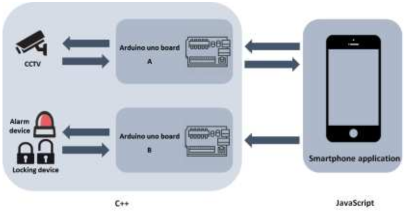


Figure 1. System Configuration Diagram

The functions of this system include a remote security function and a remote opening/closing function. The remote security function confirms the CCTV through the application, detects the suspicious person and then sounds an alarm. When there is a suspicious person, the user presses an "warning sound generation request" button. The request sent from the application is received from the door lock and the generation of the alarm sound is requested to the alarm device. The remote opening/closing function means a function for opening/closing the opening/closing device by confirming the CCTV screen by an application when opening/closing is requested. The user confirms the CCTV in the application and presses the "open/close" button when the visitor matches. Requests from the application are received from the door lock and sent to the servo motor to change the angle, and the lock opens as the servo motor changes the angle.

3 Implementation Results

This system is mounted as an Arduino board A for transmitting a CCTV video and an Arduino board B for remotely opening and closing a door. A camera module and a wireless shield are attached to the board A, and a video is transmitted to an application. A piezoelectric buzzer, a servo motor and a wireless shield are attached to the board B, and an application generates an alarm sound by utilizing the wireless communication, and opens and closes a lock device.

4 Conclusion

The project leverages WIFI communications to implement a mechanism that allows for external access, enhanced security, and open/close doors for visitors. Through this project, users will be prevented from exposing their passwords and being exposed to crime, and they will be able to reduce the risk of theft and strength when they are away from home for a long time or come home late.

Acknowledgments. This research was supported by the National Research Foundation of Korea (NRF-2018R1A4A1025559).

Improvement of TAL-based Registration in LTE Networks

Hee-Seon Jang[1], Jang Hyun Baek[2, *]

[1]Department of Computer, Pyeongtaek University
Pyeongtaek 17869, Korea
[2]Department of Industrial & Information Systems Engineering,
Chonbuk National University, Jeonju 54896, Korea
*jbaek@jbnu.ac.kr

Abstract. In long term evolution (LTE) network, the network is made up of the tracking areas (TAs) i.e. a group of cells, and several TAs make a TA list (TAL). A mobility management scheme called TAL-based registration is adopted for LTE network. In this study, we consider an improved TAL-based registration called TAL-based registration with cell-based central policy to reduce the total cost of original TAL-based registration, and obtain an accurate mathematical model to analyze the total cost of TAL-based registration with cell-based central policy. First, an improved version of TAL-based registration is explained that adopts cell-based central policy. Next, in order to accurately analyze the performance of TAL-based registration with cell-based central policy, an embedded Markov chain model based on 2-D random walk model is applied, and it is shown through mathematical analysis that the TAL-based registration with cell-based central policy is superior to distance-based registration as well as original TAL-based registration in most cases.

Keywords: 2-D random walk model; central policy; location registration; Markov chain; mobility management; TAL-based registration

In mobile cellular networks, mobility management scheme is composed of location registration (LR) and paging processes. LR is the process that enables the user equipment (UE) to register its location in network databases, whenever it enters a new location area. Paging is the process through which, when a call arrives, the network pages the UE over all cells of current location to find the cell to which the UE belongs so as to connect the incoming call to the UE. In LTE network, the mobility management entity (MME) is responsible for tracking the locations of the UEs. The MME is in charge of a group of evolved Node Bs (eNBs). The radio coverage of eNB is called as a cell. The network is composed of non-overlapped TAs i.e. a group of cells, and several TAs make a TAL. When the UE enters a TA that is not in its current TAL, the UE registers its TA to notify MME of its new TAL.

Many studies have dealt with the TAL-based registration. Chung proposed movement-based TAL forming, based on the assumption that a TA consist of a single cell. Deng et al. proposed to form the TAs as a set of rings, changing them adaptively considering the UEs' mobility. Grigoreva et al. approached signaling reduction by dynamically forming TALs according to mobility prediction and variable TAL forms. These, however, was difficult to implement and demanded for a complex paging procedure.

In TAL-based registration, every time a UE enters a new TAL, it registers its location so that the TA of entered cell is the center TA of new TAL, which is called central policy. This central policy can reduce the possibility that the UE quickly exits the new TAL. We call this type of central policy the TA-based central policy. However, this constraint makes the TAL-based registration with central policy be hard to implement in real network architecture.

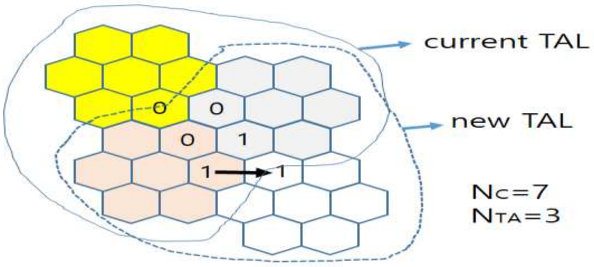


Fig. 1. An example of TAL-based registration with cell-based central policy (N_C: no. of cells in a TA, N_{TA}: no. of TAs in a TAL)

In this study, an improved TAL-based registration scheme is considered by introducing cell-based central policy to reduce total costs of original TAL-based registration. Under the cell-based central policy, the cell in which a UE registers its location is the center cell of the new TAL. For example, let's consider Fig. 1 that a TAL is composed of 3 TA which is composed of 7 cells. In this case, it is hard to say that there is a center TA and can't be applied by TA-based central policy. However, in current TAL in Fig. 1, 3 of 0-cells constitute center cells of the TAL and, if a UE exit from current TAL as marked in Fig. 1, 3 of 1-cells constitute center cells of the new TAL. Under this cell-based central policy, the TAL can have any number of TAs and as a result, it is possible to easily implement TAL-based registration with cell-based central policy in any real network architecture.

Next, we present an accurate analytical model for the proposed schemes using 2-D random walk mobility and embedded Markov chain model. When the values of any N_C and N_{TA} are given, it is possible to draw the state-transition diagram and calculate the total cost. Finally, using our analytical model, we evaluated the total cost of two types of TAL-based registration and distance-based registration. We established that, through numerical results for various circumstances, the TAL-based registration with cell-based central policy yield better performance than distance-based registration as well as original TAL-based registration.

Our study can help pick the optimal mobility management scheme and minimize the total cost on radio channels in mobile cellular networks.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2017R1E1A1A03070134) and by the Ministry of Education (2016R1D1A1B01014615).

An Efficient Image Data Crawling System Using CNN-Based Self-Training

Myoung-jae Lee[1], Young-guk Ha[1,*]

[1] Department of Computer Science & Engineering
Konkuk University
Neungdong-ro, Gwangjin-gu, Seoul 143-701, Korea
dualespresso@naver.com, ygha@konkuk.ac.kr

Abstract. As Big Data technology grew up, Artificial Intelligence technology, especially Deep Learning technology, developed rapidly. In image based Deep Learning technology, like Convolutional Neural Network, collecting accurate image data is very important for training. It can be done with image web crawler, however, collecting accurate data automatically is very difficult. This paper proposes the way for implementing accurate image crawling system that collect massive and precise image data using Convolutional Neural Network.

Keywords: Big Data, Deep Learning, Image Crawler, CNN

1 Introduction

We are living in a vast data world. In the age of vast data, research has been actively conducted to collect and analyze large amount of data and Big Data technology grew up. As the Big Data technology develops, the Artificial Intelligence, especially Deep Learning, technology that utilizes Big Data is greatly developing. To take advantage of Big Data in Deep Learning, the way to pull the right data for specific purpose is necessary. In this case, implementing a specific web crawler is recommended. However, collecting accurate data automatically is very difficult. Therefore, some additional inspection procedures are required. In this step, using CNN-based object recognition system can increase data accuracy. This paper deals with the system of an accurate and efficient image data crawling system that uses Convolutional Neural Network to collect large-scale image data.

2 Related Work

The history of crawling technology is very deep. WWW Robot is the first crawler, which was proposed in 1996[1]. They proposed a crawler concept to cover the entire

[*] Corresponding author

internet with search engines. With the advent of this concept, there has been a lot of research to improve the crawling algorithm. In 1997, Topic-specific crawling system proposed a set of themes that relatively represented the cramped portion of the web[2]. The paper on crawler similar to the current system was published in 2001, which proposed an efficient crawling system that focused on predicates on web pages and showed excellent results[3]. Another paper proposed a replication detection system for efficient crawling[4]. All of the prior systems show great performance; however, these systems need additional features to improve crawling accuracy.

3 System Design

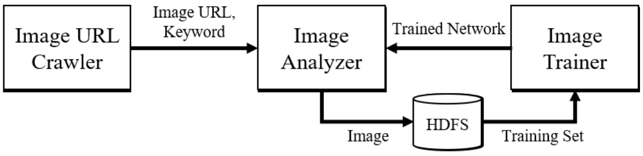


Fig 1. Overall System Architecture

As shown Figure 1 above, the proposed system consists of three core systems. Image Crawler searches images and request analysis to Image Analyzer. Image Analyzer analyses an image and determines whether the image is suitable with keyword. Image Trainer uses labeled image data stored in HDFS. It trains many training set and sends trained files, which is used to analysis images, to Image Analyzer.

3.1 Image Analyzer

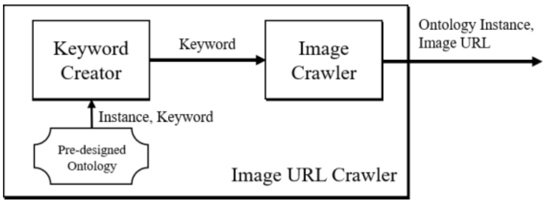


Fig 2. Image URL Crawler

Figure 2 shows system flow of the Image URL Crawler. Keyword Creator creates keyword with pre-designed ontology instance. Created keyword is send to Image Crawler and Image Crawler crawls image URL. And it sends crawled URL and ontology instance to Image Analyzer.

3.2 Image Analyzer

Image Analyzer consist of two core systems. Crawled Image Controller receives ontology instance and image URL. It sends image URL to Object Detector and receives detected result. In the Object Detector, it detects objects in the image with pre-trained network. We used framework YOLOv3 to detect objects. Pre-trained network is updated with self-trained network from Image Trainer. It means the system can become more accurate with repetitive implementation.

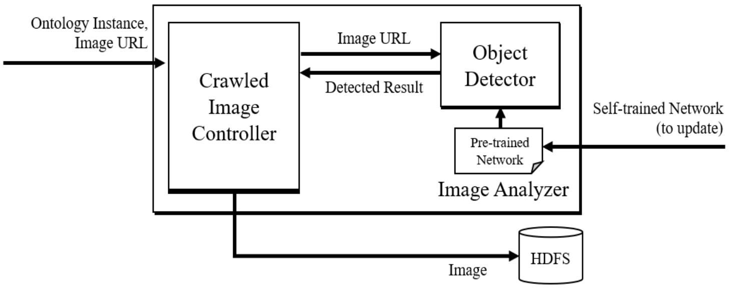


Fig 3. Image Analyzer

4 Implementation

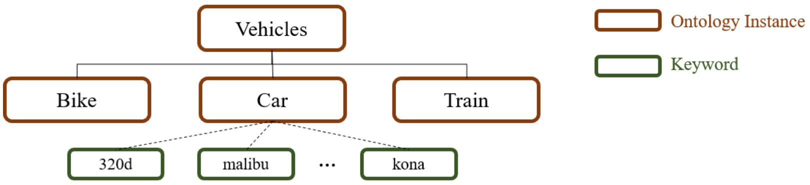
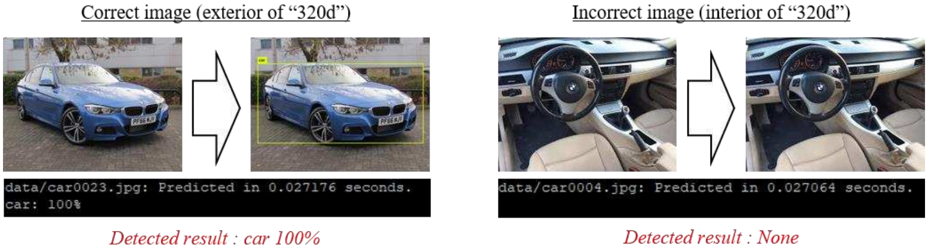


Fig 4. Pre-designed Ontology

As shown Figure 4, we designed ontology with instance and keyword. The instance will later play role as a label when recognizing the object later, and the keyword acts as a search term when crawling images. The more images there are for learning, the better. Therefore, it is better to have various keywords in one instance.


Correct image (exterior of "320d")

```
data/car0023.jpg: Predicted in 0.027176 seconds.
car: 100%
```

Detected result : car 100%

Incorrect image (interior of "320d")

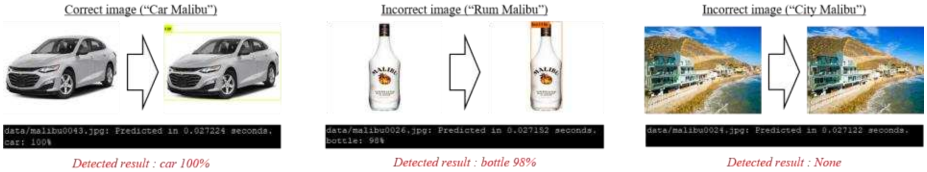
```
data/car0004.jpg: Predicted in 0.027064 seconds.
```

Detected result : None

Fig 5. The result of object detection "320d"

The implementation shows two advantages of this system. Firstly, as shown Figure 5, it can reduce images that presented in unwanted ways, although there is no problem with the meaning of words. If you search with "320d", result will not only be the exterior of the car, but also the interior. This system allows you to collect consistently in one kind.

Secondly, as shown Figure 6 below, it can eliminate the problem of ambiguous words. If you crawl for images with the keyword “malibu”, not only “Car Malibu”, but also “American coastal city Malibu” or “Rum Malibu” will be collected together. The pre-trained network in Image Analyzer of this system eliminates this problem.


Correct image (“Car Malibu”)
Incorrect image (“Rum Malibu”)
Incorrect image (“City Malibu”)

```
data/malibu0043.jpg: Predicted in 0.027224 seconds.
car: 100%
```

Detected result : car 100%

```
data/malibu0026.jpg: Predicted in 0.027152 seconds.
bottle: 98%
```

Detected result : bottle 98%

```
data/malibu0024.jpg: Predicted in 0.027122 seconds.
```

Fig 6. The result of object detection "malibu"

5 Conclusion

This paper suggested the system for implementing accurate image crawling. As system updates weights file for itself, it is evolutive system that becomes more accurate repetitive implementation. Proposed system can improve various crawling system with Deep Learning analyzer not only image crawling, but also other kind of crawler.

Acknowledgments. This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program) (20000946 , Development of artificial intelligent computing platform technology for service robots capable of real-time processing of large-capacity, high-performance sensor fusion processing and deep learning) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

References

1. Heinonen, Oskari, Kimmo Hätönen, and Mika Klemettinen. "WWW robots and search engines." (1996).
2. Chakrabarti, Soumen, Martin Van den Berg, and Byron Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." Computer Networks 31.11 (1999): 1623-1640.
3. Aggarwal, Charu C., Fatima Al-Garawi, and Philip S. Yu. "Intelligent crawling on the World Wide Web with arbitrary predicates." *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001.
4. Manku, Gurmeet Singh, Arvind Jain, and Anish Das Sarma. "Detecting near-duplicates for web crawling." *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007.
5. Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

A Comparative Analysis of Representative Learning Characters for Deep Learning Based Hangul Automatic Generation

Dong-Yeon Park[1], Young-Seo Ji[2], Soon-Bum Lim[3]

[1,2] IT Engineering, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 04310, Korea
[3] Research Institute of ICT Convergence, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 04310, Korea
yeon0729@sookmyung.ac.kr, 1713786@sookmyung.ac.kr, sblim@sm.ac.kr

Abstract: Recently, font creation technology based on deep learning has been increasing. However, the result is different according to the selection of representative characters. In order to find the characters that produce the best results in automatic generation of Hangul characters, we divide the composition principle of Hangul into 3, 6, 9 kinds and derive optimal construction principle in automatic character generation based on deep learning using GANJY FONT.

Keywords: Hangul Composition Principle, Deep Learning, Automatic Character Generation

1 Introduction

In the new process of creating Hangul fonts and characters, many characters need to be tested several times in order to grasp the readability and design of the characters and correct them according to the problems. However, it is impossible to test using such a few hundred characters in the Hangul character automatic generation service.

The purpose of this study is to construct a set of characters within a few hundreds of characters which enable quick identification and verification of readability and design in the Hangul character automatic generation service. Here, the automatic generation of Hangul characters uses the GANJY FONT[1] service which automatically generates Hangul characters based on the deep learning.

2 Main Subject

2.1 Outline of research process

The research proceeds in the following order. Research plan and composition Principle design, creation of the evaluator format, collection of the evaluator font, scanning and correction of the collection data, construction of the GANJY FONT[1] learning model, extraction of the collected font data, evaluation of the evaluator format, evaluation and analysis of the result data.

2.2 Designing composition principle of Hangul

The composition principle of Hangul is divided into 3, 6, and 9 kinds. First is a character structure without both a final consonant and a double consonants. Second is a character structure with final consonant but no double consonants. Last is a character structure with both a final consonant and a double consonants.

The sample characters that represent each composition principle are as follows. Therefore, a set of characters to be given to the appraiser is composed of 31 characters in total. As a result, 310 characters are collected for 10 appraisee

Table 1 Learning and Result Sentence based on Hangul composition principle

Composition Principle	Learning Sentence	Final consonant	Double consonant
3	수채화	X	X
6	웹시각화표준	O	X
9	밝고 환한 꽃의 잎을 땄다	O	O
Result Sentence			
꿈꿨는지 말해줘요 원더걸스			

2.3 Establish evaluation criteria

Evaluation criteria are divided into completeness and similarity. Completeness refers to the meaning of a character, and specifically evaluates whether the character is properly generated to be recognizable. Similarity refers to the style of a character, which evaluates whether the style and font of the resulting character are similar to the original author's font.

Acknowledgments. This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No. 2016-0-00166)

References

1. Ja-Yeon Jeon, Joo-Young Yang, Soon-Bum Lim, 「Implementation of Automatic Hangul Handwriting Production Service Using Deep Learning」, 『한국 HCI 학회 학술대회』, 한국 HCI 학회, 2019.

Improved Paging Scheme of Distance-based Registration in Mobile Communication Networks

Gwang Jin Han[1], Hee-Seon Jang[2*]

[1]Department of Industrial & Information Systems Engineering
Chonbuk National University, Jeonju, 54896, Korea
[2]Department of Convergence Software, Pyeongtaek University
Pyeongtaek, 17869, Korea
[*]hsjang@ptu.ac.kr

Abstract. To reduce the signaling cost occurring in the boundary cells of registration area (RA), 2-location distance-based registration (2DBR) scheme was previously proposed. The 2-step selective paging algorithm was also adopted in the 2DBR, and its performance was compared with 1-location DBR with 1-step simultaneous paging. For the fair comparison, we propose the ring-based and cell-based paging scheme in the traditional DBR. Through the simulation studies, it is observed that the performance of the paging schemes depends the call-to-mobility ratio, distance threshold, unit registration and paging cost, and so on.

Keywords: location registration, paging, distance-based location registration (DBR), 2DBR, mobility management.

In mobile networks, it is important to know the location of mobile station (MS) so that the incoming calls can be connected. The network uses two essential operations (location registration and paging) to keep track of the MS. In general, there exists a trade-off relation to treat the signaling traffic between the location registration and paging. So, until now, many researches have been studied to reduce the signaling cost for the registration and paging.

Among them, the distance-based registration (DBR) scheme provides relatively good performance as compared with other registration algorithms as much as the MS resides in the center cell the MS registers last. In the DBR, an MS will register when the distance between the current cell and the cell in which registration last occurred exceeds a prescribed distance threshold *d*. To avoid the registration requests, Baek *et al*. proposed 2-location distance-based registration (2DBR). The 2DBR stores not only the RA that an MS registered last but also the RA that an MS registered second to last, and prevent the MS from re-registering anew when it alternates between stored RAs. And, for the paging scheme to connect incoming call of MS, they also proposed the 2-step selective paging algorithm as follows.

- All the cells of new RA are paged first.
- The rest of the cells are paged next if there is no response.

They also provided the numerical results for the DBR and 2DBR with 2-step selective paging algorithm. However, note that the 1-step paging scheme was adopted in the DBR for comparison with 2DBR. That is, in the DBR, all cells in the RA were paged simultaneously. In this paper, we propose the 2-step selective paging algorithm in the DBR, and compare the DBR and 2DBR under the same environments of paging strategy for the fairness.

For example, assuming hexagon cell configuration and *d*=3, Fig .1 shows the RA lists for each scheme when the MS moves A→B. In the 2DBR, to find the MS, new RAs are paged (cells 3, 4, 10~13, 20~32), and then the rest of cells (cells 1, 2, 5~9, 14~19) are paged if no response. In the DBR with one RA, we propose two different paging scheme as followings.

- Ring-based paging (RBP): For example, ring 1 and 2 are paged simultaneously, and ring 3 is then paged if there is no response from the MS.
- Cell-based paging (CBP): For example, ring 1 (cell 20) and 4 cells of ring 2 (cells 21~24) are paged simultaneously, and remaining cells (cells 11, 12, and all of ring 3) is then paged if there is no response from the MS.

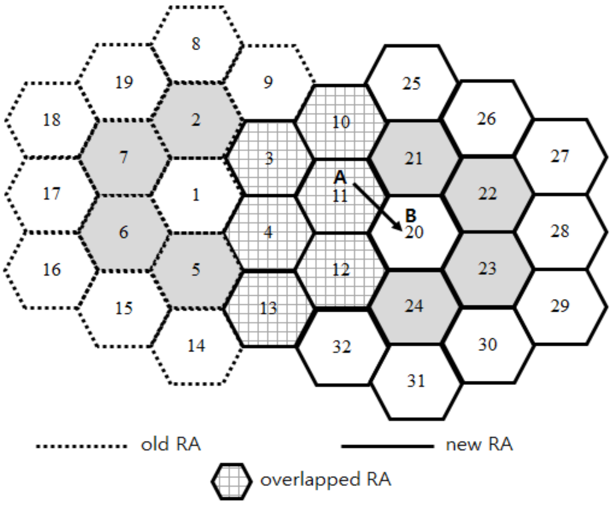


Fig. 1. Distance-based registration with two RAs (*d*=3)

Note that as the unit paging cost has a higher value, the performance of the registration algorithm depends on the paging strategy and the probability that the MS resides in each cell. For further study, the simulation studies for the DBR and 2DBR with various paging schemes will be performed. The performance for two paging schemes will be also compared. Our results can help to decide the optimal paging scheme to minimize the signaling cost on radio channels for distance-based registration location management. The detailed numerical results will be explained in the presentations.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (Grant Number 2017R1E1A1A03070134).

Improved Spatial Modeling using Path Distance Metric for Urban Traffic Prediction*

Sung-Soo Kim[1,2], Okgee Min[1], and Young-Kuk Kim[2]

[1] Smart Data Research Group, ETRI, Daejeon, South Korea
{sungsoo, ogmin}@etri.re.kr
[2] Chungnam National University, Daejeon, South Korea
ykim@cnu.ac.kr

Abstract. Traffic prediction plays a crucial role in reducing traffic congestions and improving transportation. However, the traffic prediction problem is challenging due to complex spatial dependency and temporal dynamics, which are difficult to model. To represent accurate spatial properties of the road network, we propose the weight modeling technique for the adjacency matrix using the *path distance metric* for the graph signals. The experimental result shows that the recent deep learning techniques with the proposed spatial model are promising solutions to the traffic prediction in terms of accuracy.

Keywords: Traffic Prediction · Spatio-Temporal Modeling · Graph Embedding.

1 Introduction

Traditionally, many studies in traffic prediction problems have exploited autoregressive integrated moving average (ARIMA) and Kalman filtering [1] to predict the future traffic speeds in the road networks. *Data-driven prediction approaches* using deep learning models for traffic forecasting have received wide attention for recent years [2][3].

Motivation: To provide traffic forecasting, spatio-temporal modeling of the road network is one of the important tasks. Previous studies exploited Euclidean distance with thresholded Gaussian kernel to compute the weights in the adjacency matrix [2][3]. These methods used spatio-temporal modeling, but do not represent accurate spatial dependency according to the real-world road networks as shown in Figure 1-(a).

Problem Definition: *Traffic Prediction Problem.* Given the linear dual graph G for the road network $\mathcal{R}$ and the historical traffic time series $\mathcal{T}$ until time interval t, the traffic speed prediction problem aims to learn a function $\mathcal{P}(\cdot)$ for predicting the future traffic speeds $\hat{\mathcal{T}} = \{\mathbf{X}^{(t+1)}, ..., \mathbf{X}^{(t+T)}\}$ for each link in $\mathcal{R}$. $\mathbf{X}^{(t)}$ denotes the graph signal observed at time t.

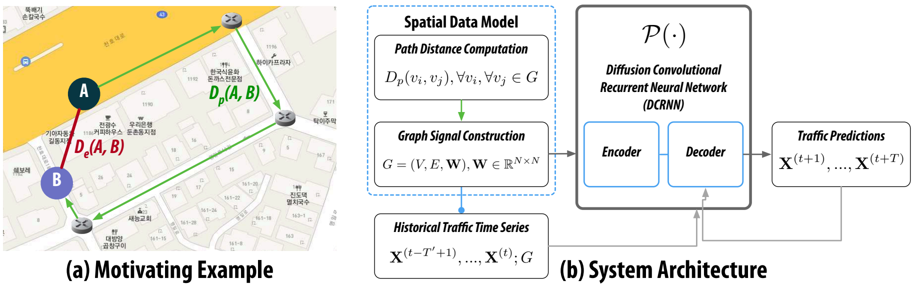
$$\Big[\mathcal{T} = \{\mathbf{X}^{(t-T'+1)}, ..., \mathbf{X}^{(t)}\}; G\Big] \rightarrow \mathcal{P}(\cdot) \rightarrow \Big[\mathbf{X}^{(t+1)}, ..., \mathbf{X}^{(t+T)}\Big] \qquad (1)$$

The edge weights $w_{ij} = \exp(\frac{-kD_p^2(i,j)}{\sigma^2}), i \neq j$ otherwise 0, denotes pairwise *proximity* in the weighted adjacency matrix $\mathbf{W}$, where k is the number of *hops* from i to j and σ is the standard deviation of path distances. Our main contribution is that we present the improved *nodes proximity* representation using the *path distance* metric with the graph topological features such as k-hop and node degrees for the urban road network.

* **Acknowledgements.** This work was supported by Institute of Information & Communication Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00121, *Development of a Traffic Predictive Simulation SW for Improving the Urban Traffic Congestion*). Also, the authors would like to thank Jae-Wook Lee for helpful discussions and comments related to the preparation of training dataset for traffic prediction.

2 System Architecture

Our system architecture contains two major parts such as the spatio-temporal traffic data model and the deep learning model as shown in Figure 1-(b).


A
B
$D_p(A, B)$
$D_e(A, B)$
Spatial Data Model
Path Distance Computation
$D_p(v_i, v_j), \forall v_i, \forall v_j \in G$
Graph Signal Construction
$G = (V, E, \mathbf{W}), \mathbf{W} \in \mathbb{R}^{N \times N}$
$\mathcal{P}(\cdot)$
Diffusion Convolutional Recurrent Neural Network (DCRNN)
Encoder
Decoder
Traffic Predictions
$\mathbf{X}^{(t+1)}, ..., \mathbf{X}^{(t+T)}$
Historical Traffic Time Series
$\mathbf{X}^{(t-T'+1)}, ..., \mathbf{X}^{(t)}; G$

(a) Motivating Example (b) System Architecture

Fig. 1. Motivating example and system architecture: (a) A and B are links in the road network. The *path distance* $D_p(A, B)$ more accurately represents the real-world urban road network than *Euclidean distance* $D_e(A, B)$. $\mathbf{W}$ denotes a weighted adjacency matrix ($N = |V|$). (b) The proposed system exploits D_p metric for spatial modeling. Then, the spatial data and the historical traffic time series are fed into the encoder and the decoder of the traffic prediction model $\mathcal{P}(\cdot)$.

First, we construct the *linear dual graph* from the primal graph of the road network. Then we can represent the graph as a weighted directed graph $G = (V, E, \mathbf{W})$, where V and E are a set of nodes and a set of edges, respectively. To represent accurate nodes proximity, we use the *path distance metric* D_p rather than the pairwise Euclidean distance D_e. Our system performs *all pairs shortest path* algorithm to compute the path distance. We use the diffusion convolutional recurrent neural network (DCRNN) as a deep learning model. The DCRNN generalizes convolution to graphs based on the *propagation* characteristics of traffic and learns a representation for each node [2].

Experiments: For experiments, we used the road network for Cheonho-daero of Seoul in South Korea. The historical traffic time series has 7 month-range traffic data. We exploit the mean absolute percentage errors (MAPE), $M(t, p) = \frac{100}{n} \sum_{i=1}^{n} |\frac{t_i - p_i}{t_i}|$, to measure the performance (t: ground truth, p = predicted speeds). The MAPE result of preliminary experiment is 6.0% for ahead traffic prediction of 80 links during 24 hours from 15 minute-forecasting horizon, which is better than the result (6.13%) using D_e.

3 Conclusion

We have proposed the weight modeling technique using *path distance metric* for spatial modeling of the road network. Our approach reflects the accurate spatial correlation between links according to the urban road network. This work showed that deep learning techniques with our spatial model are promising solutions to traffic prediction.

References

1. Kim, S.S., Kang, Y.B.: Congestion Avoidance Algorithm Using Extended Kalman Filter. In: 2007 International Conference on Convergence Information Technology (ICCIT 2007). pp. 913–918 (Nov 2007)
2. Li, Y., Yu, R., Shahabi, C., Liu, Y.: Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In: 2018 International Conference on Learning Representations (ICLR 2018) (2018)
3. Yu, B., Yin, H., Zhu, Z.: Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In: 2018 International Joint Conference on Artificial Intelligence (IJCAI). pp. 3634–3640 (2018)

Air Quality Prediction using Meteorological Data based on Data Mining Model

Menghok Heak[1], Sang-Hyun Choi[1],

[1] Department of Manangement Information System, Chungbuk National University,
Cheongju, South Korea
menghok.heak@gmail.com, chois@cbnu.ac.kr

Abstract. Air pollution is one of the most serious issues to many countries. This environment issue causes a lot of harmful diseases to human health in physical and mental. In Korea, PM10 is one of the most concerning issues in air pollution. Meteorological condition is an impacting factor of this issue. This research will study about the prediction the air quality index in Korea focusing on PM10 by using the meteorological data as the predictors. The paper will discuss about the previous research of air pollution in section 2, the characteristic of data for this study in section 3 and results of the study in section 4.

Keywords: Air Pollution; Prediction; PM10;

1 Introduction

Air pollution is one of the most serious issues for any countries around the world. This environmental issue causes a lot of harmful diseases to human health both to the physical and mental illness [1]. It was estimated that, around 7 million deaths annually were caused by air pollution, approximately 1/8 of premature death [2]. The root cause of the pollution is believed to be from the industrial emissions, vehicle engines generations and meteorological factors [3][4]

The index of air quality is indicated and calculated using methods and measurements differently depending on the specific region or country. However, the common measurements among all are Atmospheric Particle Matters or Particle Matters (PMs) [5]. The degree of harmful health effects of air pollution depends on the size of the particles. PM10 and PM2.5 are defined regarding to the diameters of the particles. PM10 refers to the particle with the size of its diameter is between 2.5µm and 10µm, so called fine dust, and PM2.5, so called ultrafine dust, is the particle with a diameter of 2.5µm or less.

Particularly in Korea, one of the concerning air pollution indices is PM10. This research is focusing on the predicting the air quality index considering on the meteorological information.

2 Related Researches

Beside the human's activities factors such as manufacturing productivities or vehicle usages that produce the pollutant, meteorological condition is one of the most impacting factors influencing accumulation and elimination of the air pollution issue [6]. The pollutant produced in another region may move to another region through the wind flow. Wind speed, wind direction, temperature, humidity, and other weather relating measurements also join in the producing and transferring process of the pollutant as well. The issues of air pollution usually occur during winter and summer which have the extremely low and high temperature [7].

It is believed that, a part of the pollution in Korea is causing by the manufacturing production in neighbor country which is delivered to Korea through the wind flow [6]. Air quality index in Korea considers on the 6 measuring components – PM10, PM2.5, Ozone (O3), Nitrogen Dioxide (NO2), Carbon Monoxide (CO) and Sulfur dioxide (SO2) [8]. However, one of the most concerning components is PM10. Many parts of Korea got suffered from the cause of PM10. The degree of PM10 in Korea is divided in the 4 levels – 0-30, 31-80, 81-150 and greater than 150 to identify the seriousness of its effects where the lowest number cause least harmful [8].

3 Data Characteristic

Data used for this research was captured from Korea's Government open data platform and Korean Ministry of Environment's AirKorea site between January 1st, 2013 and December 31st, 2018. The data is hourly interval. By using the division rule from AirKorea, the data can be divided in to 4 levels.

Table 1. The degree of PM10 was categorized into 4 levels as suggested by AirKorea.

Level Name	Degree of PM10
Good (G)	0 – 30
Moderate (M)	31 – 80
Unhealthy (U)	81 – 150
Very Unhealthy (VU)	150 +

The dataset consists of 8 dimensions – Datetime, Temperature, Humidity, Windspeed, Wind Direction, Snow Depth, Raining Condition and PM10. The level of PM10 is not only limited to the current meteorological condition. For instance, the level of PM10 usually drop 1 hour after raining not at the time of raining started. Therefore, using previous record of data is required for the analysis and prediction.

4 Results

As the results of the research, built model is acceptable for the predicting of this meteorological data. As shown in the figure 1 of the AUC-ROC Curve, the least accurate prediction is “Moderate” level which has value of 0.88 (88%) accurate, while the highest one is “Very Unhealthy” level with 0.97 (97%). The accuracy for the “Good” level is 0.92 (92%) and 0.95 (95%) for “Unhealthy” level. Averagely, the model is performing with the accuracy of 0.93 (93%) which is highly accurate.

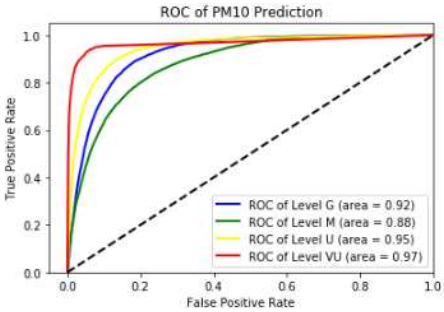


Fig. 1. AUC-ROC Curves Graph of PM10 Prediction

5 Conclusion and Future Research

The result is telling that the model with accuracy of 93% is acceptable and can be used for predicting the level of PM10 by using the meteorological information and previous degree of the pollutant. However, the model is suitable only for short-term prediction.

For recommendation, as the data is in time series with hourly interval, the analysis using time-series data analysis algorithm is recommended for further research. By using the time-series data analysis algorithm, the new model would be able to predict in longer-term.

References

1. Pope, C.A., Burnett, R.T., Thun, M.J., et al: Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution. Jama-J. Am. Med. Assoc. 2002, 287, 1132–1141 (2002)
2. WHO: Air Pollution, Climate and Health. Retrieved on June 28th, 2019 from https://www.who.int/sustainable-development/AirPollution_Climate_Health_Factsheet.pdf
3. Chang Z., Guojun S.: Application of data mining to the analysis of meteorological data for air quality prediction: A case study in Shenyang. doi :10.1088/1755-1315/81/1/012097 (2017)
4. Zhao C. X., Wang Y. Q., Wang Y. J., Zhang H. L., Zhao B. Q.: Temporal and Spatial Distribution of PM2.5 and PM10 Pollution Status and the Correlation of Particulate Matters and Meteorological Factors During Winter and Spring in Beijing, Environmental Science, 2014, 35(2):418-427, (2014)

5. WikiPedia: Air Quality Index. Retrieved on June 28th, 2019 from https://en.wikipedia.org/wiki/Air_quality_index
6. Soonae P., Hyunjae S.: Analysis of the Factors Influencing PM2.5 in Korea: Focusing on Seasonal Factors. http://dx.doi.org/10.15301/jepa.2017.25.1.227 (2017)
7. German H., Terri A. B., Shannon L. W., David P.: Temperature and Humidity Effects on Particulate Matter Concentrations in a Sub-Tropical Climate During Winter. International Proceedings of Chemical, Biological and Environmental Engineering (2017)
8. Air Korea Website. Retrieved June 28th, 2019 from https://www.airkorea.or.kr/index

Real-Time Accessibility Evaluation Technique for EPUB Document

Hyeongki Cho[1,2], Jin-Suk Kim[1], Kwan-Hee Yoo[2*]
[1]NeoForce Co., Ltd., South Korea
gudrl0517@gmail.com
[2]Dept. of Computer Science, Chungbuk National University, South Korea
khyoo@chungbuk.ac.kr
*Corresponding Author

Abstract. EPUB is designed on web-based environment and is eventually optimized for smart devices. According to extension of usage of the EPUB, it is suitable for making e-books for the non-disabled. Even by the early EPUB standard was established without consideration of accessibility for the non-disabled, it has evolved into a format that supports both the non-disabled and disabled since EPUB 3.x. Specially, DAISY and HTML5 accessibility technology is adapted into the EPUB. Therefore, there it is necessary to study techniques for evaluating EPUB accessibility. The paper proposes an evaluation system for the accessibility of EPUB document. The proposed evaluation system uses a lightweight evaluation technique based on the SAX parser so that the evaluation can be performed in real time.

Keywords: EPUB, Accessibility, ACE

1 Introduction

Digital Accessible Information SYstem (DAISY) [1] and Electronic Publishing (EPUB) [2] are a typical e-book format used as a substitute for reading disabled people. DAISY is an audiobook format for the reading disabled people. And many alternatives have been made in this format. But this format is not suitable for non-disabled people who want to read in general, because it is a format focused on audiobooks. EPUB, on the other hand, is optimized for smart devices and it can be designed with a web-based, so it is suitable for making e-books for the non-disabled. The early EPUB standard was established without consideration of accessibility. But since EPUB 3, it has evolved into a format that supports both non-disabled and disabled people by accepting DAISY technology and HTML5 accessibility technology [3].

This study proposes an evaluation system and technique for the accessibility of EPUB e-books. The proposed evaluation system uses a lightweight evaluation technique based on the SAX parser. And that allows the system to evaluate in real time.

2 Related Work

Currently, a representative tool for evaluating the accessibility of EPUB e-books is the Accessibility Checker for EPUB (ACE) [4] published by the DAISY consortium. ACE is based on aXe, an open source web accessibility evaluation tool of Deque Systems Inc. Therefore, 59 out of 64 evaluation items are related to the Web. This means that the ACE does not fully support the specific evaluation items to the EPUB format. In addition, ACE renders web content using an embedded web browser, so its evaluation process is very slow.

Another case study is the implementation of an EPUB accessibility verification component that follows the structure of EPUBCheck from IDPF and is based on the TTA eBook Accessibility Guide [5]. The study proposed 48 verification items and used commercial e-books of Korean publishers as samples to test them. However, since the sample mostly consisted of simple text and pictures, the verification results tended to be biased toward text or style related items. This study presents more automated evaluation items than other studies' by closely analyzing EPUB Accessibility 1.0 and TTA Accessibility Guidelines. In addition, based on these EPUB accessibility standards, we propose evaluation items that are more specific to EPUB accessibility than ACE. In particular, to improve the very slow evaluation speed of ACE, this paper proposes a faster and lighter evaluation algorithm that can replace the rendering process.

This study used 45 EPUB 3 samples provided by IDPF to experiment with the proposed evaluation items and algorithms. The EPUB 3 sample includes examples of various features of EPUB 3, such as media overlay and fixed layout. So, they are very suitable for testing all evaluation items and algorithms.

3 Evaluation Technique

Table 1. Evaluation Target Document and File Form

Target Document	Target File Type	Target File Format
Package Document	OPF	XML
Content Document	XHTML	XML
	SVG	XML
	CSS	CSS
Media Overlay Document	SMIL	XML

As shown in Table 1, all documents to be evaluated using XML format except CSS. Therefore, the proposed evaluation technique uses SAX parser to efficiently access XML format. CSS uses CSS parsers of EPubCheck, which has similar structures to SAX parsers.

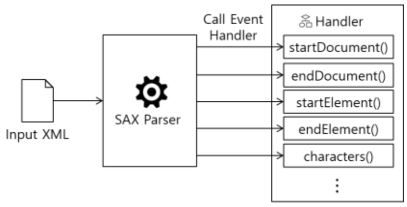


Fig. 1. Structure of SAX Parser

As shown in Figure 1, the Simple API for XML (SAX) parser consists of handlers that handle events for specific statements. Since it has this structure, the algorithm must be defined on a per-event. In addition, all the algorithms have a similar flow and are divided into six types according to the inspection range as shown in Table 2.

Table 2. Evaluation Target Document and File Form

Inspection Scope	Description	Algorithm Type
Element	Very simple inspection scope to check only the elements to be inspected	**Type1.** Type specific to element scope **Type2.** CSS inspection type
Document	Scope that references the parent, child, or neighboring element of the element to check within a single document	**Type3.** Types where references occur between elements that have a containment relationship **Type4.** Types that need to inspect the entire document
Container	Scope of algorithms that require references between elements of different documents within a container	**Type5.** Simple inspection type on container category **Type6.** Types of algorithms that require complex references in container scope

Type 1 includes the most evaluation algorithms. It is evaluated using only the element's appearance or its attribute value.

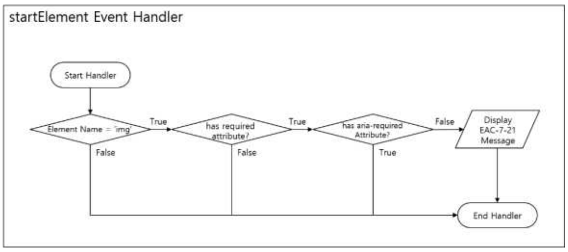


Fig. 2. Algorithm to check whether using 'aria-required' attribute

The algorithm in Figure 2 checks if the 'aria-required' attribute was also specified when the 'required' attribute was used on the 'input' element. The algorithm only refers to the tag name and attribute value of an element.

4 Performance Analysis

This study compared the processing time of ACE and the proposed evaluation technique for 45 EPUB 3 samples provided by IDPF. For accurate experimental results, each sample was tested 10 times and compared with the average of the processing time.

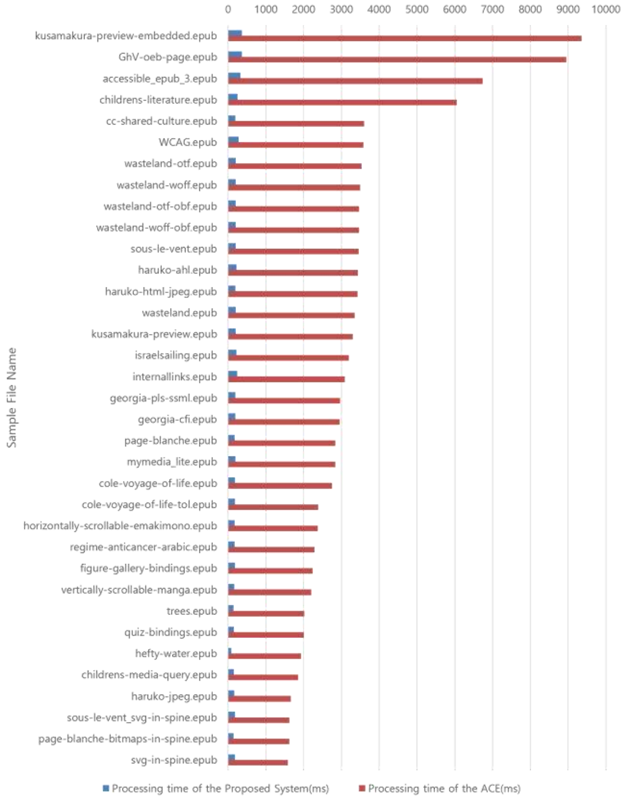


Fig. 3. Comparison of Processing Time between Proposed Technique and ACE

Figure 3 shows the difference in processing time between the ACE and the proposed evaluation technique for samples except the top 10 samples. The top 10 samples were excluded because the differences were too large to be graphically represented.

As can be seen from the graph, the sample with the largest difference in processing time showed 25 times difference, and the other samples also show a big difference. The graph in Figure 4 shows that processing time tends to be affected by the amount of web content. This can be seen that rendering has a big impact on processing time.

5 Conclusion

In terms of processing speed, the proposed method is superior to ACE, but the quality of the inspection items is lower than that of ACE, which renders web content, especially in the style area. Therefore, it is necessary to study a lightweight rendering engine with fast processing speed and an evaluation technique using this engine.

References

1. ANSI/NISO: Specifications for the Digital Talking Book [DAISY 3]. Z39.86-2005 (R2012), (2005).
2. EPUB 3 Overview, https://www.idpf.org/epub/30/spec/epub30-overview.html.
3. Kang S.-Y., Lim K.-W.: A Study on the Change of Production Strategy and Environment for Alternative Material for Person with Special Needs. Journal of Korean Library and Information Science Society. Vol. 48(4), p.283-301 (2017).
4. An accessibility checker for EPUB, https://daisy.github.io/ace/.
5. Kim, H.-Y., Lim, S.-B.: Accessibility Automatic Inspector Library for EPUB and its Components. Journal of Korea Multimedia Society. Vol. 20, pp.330–335 (2017).

IDNet : Inception-Based Densely Convolutional Neural Networks

Cheol-jin Kim[1], Young-guk Ha[1, *]

[1] Department of Computer Science & Engineering,
Konkuk University.
120 Neungdong-ro, Gwangjin-gu, Seoul, Korea
cjfwls1070@naver.com, ygha@konkuk.ac.kr

Abstract. With the explosive interest and research of deep learning, convolutional networks have become an integral part of the computer vision. Recent work has shown a variety of network models with excellent performance for computer vision. In this paper, we introduce IDNet, which finds and combines complementary points between two impressive and powerful networks (inception-family and denseNet).

Keywords: Deep neural network, Neural network architecture, Computer vision, Depthwise convolutional layer, Channelwise convolutional layer.

1 Introduction

Advances in computer hardware (particularly GPUs) and the study of various network architectures have shown high performance in machine learning technology in various fields. In computer vision, convolutional neural networks (CNNs) have become the dominant presence. As evidence, solutions using CNNs in most areas of computer vision (image classification, semantic segmentation, image captioning, etc) show the highest performance.

2 Related Work

Most networks have designed their networks deeply in order to achieve good performance. Looking at previous studies that are key to image classification, network have evolved to deeper and deeper (5-layer LeNet[1], 8-layer AlexNet[2], 19-layer VGGNet[3] and 100-layer ResNet[4]).

Paradoxically, simply making the network deeper results in poor performance. The reason for this is that as the information of the input and the gradient goes deeper into the layer, the values become dull (commonly called vanishing-gradient problem).

[*] Corresponding author

Many recent studies have seriously addressed these issues and tried to alleviate this problem even as networks deepen. ResNet[4] introduced the concept of a Skip Connection, which allows for better forwarding of input information and gradient even in deep networks. DenseNet[5] dramatically alleviated this problem by applying Dense Connectivity.

There is also a study on GoogLeNet[6], which has a deep network structure and performs well using various scale filter. In particular, GoogLeNet[6] extracts features of various scales through the Inception Module, but does not significantly increase the amount of computation. This effectively increased the network size, resulting in outstanding performance at ILSVRC2014. In subsequent studies of the Inception Module, various versions of the Inception Module were released [7],[8],[9]. Especially, Xception[9] composes the Inception Module in a form similar to Depthwise Separable Convolution. It is a completely separated structure of cross-channel correlation and spatial correlation, obtain better performance.

3 IDNet

IDNet is a network that finds and combines the complementarities of DensNet[5] and Inception Family[6],[7],[8],[9]. We propose two types (A type, B type) network architecture by applying key idea of counterpart network to each network.

3.1 A type (base DenseNet + Sparse Architecture)

For DenseNet[5], only 3x3 Conv is used in the Dense Block. We use 3x3 Conv, 5x5 Conv, and 7x7 Conv for DenseBlock, focusing on how to maintain a sparse structure while extracting various feature with multiple scale filters of the Inception Module.

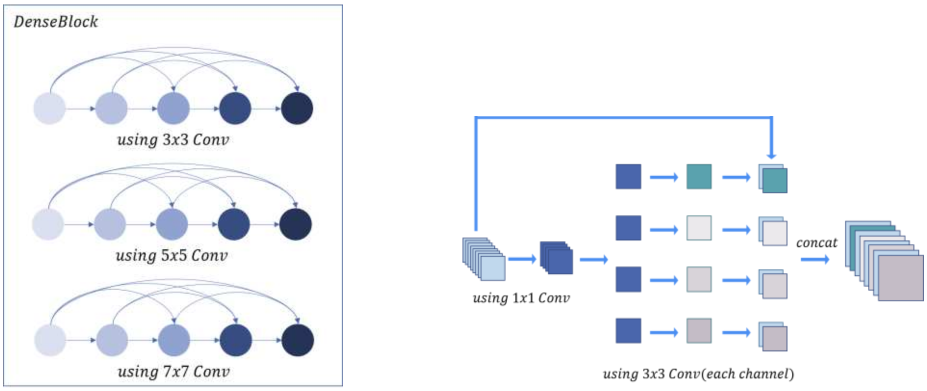


Fig. 1. Left is a DenseBlock of A type with different scale filters, and Right is a depthwise convolutional process of B_2 version with Dense Connectivity.

3.2 B type (base Xception + Dense Connectivity)

B type is composed of 4 versions. In the architecture of Xception[9], the concept of Skip Connection in ResNet[4] is used. The B_1 version configures this part by changing it to Dense Connectivity. For other versions, delete the Skip Connection part of Xception[9]. The B_2 version uses Dense Connectivity only for the depthwise convolutional parts. In contrast, the B_3 version uses Dense Connectivity only for channelwise convolutional parts. Finally, the B_4 version uses Dense Connectivity for both depthwise and channelwise parts.

4 Conclusion

We analyzed two very important and powerful networks in the field of computer vision, especially in image classification. And we thought that the two networks could be improved complementary, and based on this, we propose IDNet. Future work will use various datasets to verify the performance of IDNet, and to tune network structures and set up hyperparameters to achieve high performance with moderate increases in computation.

Acknowledgments. This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. 2018R1A2B2006154)

References

1. Yann LeCun, Leon Bottou, Yoshua Bengio, Patrick Haffner.: Gradient-Based Learning Applied to Document Recognition. IEEE (1998)
2. Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton.: ImageNet Classification with Deep Convolutional Neural Networks, NIPS (2012)
3. Karen Simonyan, Andrew Zisserman.: Very Deep Convolutional Networks for Large-Scale Image Recognition. ICML (2015)
4. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.: Deep Residual Learning for Image Recognition. CVPR (2016)
5. Gao Huang, Zhuang Liu, Laurens van der Maaten.: Densely Connected Convolutional Networks. CVPR (2017)
6. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich.: Going Deeper with Convolutions. CVPR (2015)
7. C. Szegedy, V. Vanhoucke, S. loffe, J. Shlens, Z. Wojna.: Rethinking the Inception Architecture for Computer Vision. CVPR (2016)
8. C. Szegedy, S. Ioffe, V. Vanhoucke, Alexander A. Alemi.: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. AAAI (2017)
9. Francois Chollet.: Xception : Deep Learning with Depthwise Separable Convolutions. CVPR (2017)

MULTIMEDIA

LIFE

STORAGE

NETWORK

DATABASE

SYSTEM

BIG DATA

SCIENCE

CLOUD

TREND

CLUSTER

ANALYSIS

BUSINESS

SOCIETY

GRAPHICS

VISUALIZATION

THE KOREA
BIG DATA SERVICE SOCIETY
한국빅데이터서비스학회

N13 404-2, Chungbuk University, Chungdae-ro 1
Seowon-Gu, Cheongju, Chungbuk 28644, Korea

Tel: 043-261-3636 / Fax: 043-266-3637

Email: kbigdataservice@gmail.com

Homepage: www.kbigdata.or.kr