https://doi.org/10.5392/IJoC.2019.15.4.113

Relations Between Paprika Consumption and Unstructured Big Data, and Paprika Consumption Prediction

Yongbeen Cho
Agricultural Bigdata Division
Rural Development Administration, Jeonju-si, 54875, South Korea

Eunhwa Oh
Department of Big Data
Chungbuk National University, Cheongju 28644, South Korea

Wan-Sup Cho
Department of Management Information Systems
Chungbuk National University, Cheongju 28644, South Korea

Aziz Nasridinov, Kwan-Hee Yoo
Department of Computer Science
Chungbuk National University, Cheongju 28644, South Korea

HyungChul Rah
Department of Big Data Convergence
Chungbuk National University, Cheongju 28644, South Korea

ABSTRACT

It has been reported that large amounts of information on agri-foods were delivered to consumers through television and social networks, and the information may influence consumers' behavior. The purpose of this paper was first to analyze relations of social network service and broadcasting program on paprika consumption in the aspect of amounts to purchase and identify potential factors that can promote paprika consumption; second, to develop prediction models of paprika consumption by using structured and unstructured big data. By using data 2010-2017, cross-correlation and time-series prediction algorithms (autoregressive exogenous model and vector error correction model), statistically significant correlations between paprika consumption and television programs/shows and blogs mentioning paprika and diet were identified with lagged times. When paprika and diet related data were added for prediction, these data improved the model predictability. This is the first report to predict paprika consumption by using structured and unstructured data.

Key words*: Agri-food, Prediction, Unstructured Big Data, Paprika.*

1. INTRODUCTION

Recently, we reported on prediction of onion purchase using structured and unstructured big data in order to analyze impacts of social network service (SNS) and broadcasting program on onion purchase and identify potential factors that can promote onion consumption when agricultural products face over-production and prices fall subsequently [1]. Our paper was based on the reports that information on agri-food was delivered to the consumers through mass media including television and social networks, and the information may affect consumers' behavior [2]-[4]. Recent studies report that prediction of economic activities by using social network data or internet search data ahead actual activities in stock market,

* *Corresponding author, Email: hrah@cbnu.ac.kr*
Manuscript received Dec. 20, 2019; revised Dec. 30, 2019; accepted Dec. 30, 2019

marketing and tourism [5]-[7] as well as agriculture [4], [8]-[11].

In this paper, we aimed to analyze relations of SNS and broadcasting program on paprika consumption in the aspect of amounts to purchase paprika and identify potential factors that can promote paprika consumption in advance. We also aimed to develop prediction models of paprika consumption by using structured and unstructured big data.

According to the news reports in Korea, recent increase of paprika cultivation areas and increased annual outputs due to its high price as well as prolonged economy recession led to fall in paprika consumption and paprika price as seen in Figure 1 [12]. To data, paprika consumption prediction that used structured and unstructured data has never been reported. The method used in our previous report on onion purchase prediction was applied with modification in order to identify potential relations between paprika consumption and news, broadcasting programs and SNS [1]. The relations between paprika consumption and news, broadcasting programs and SNS were applied to improve the prediction model of paprika consumption.

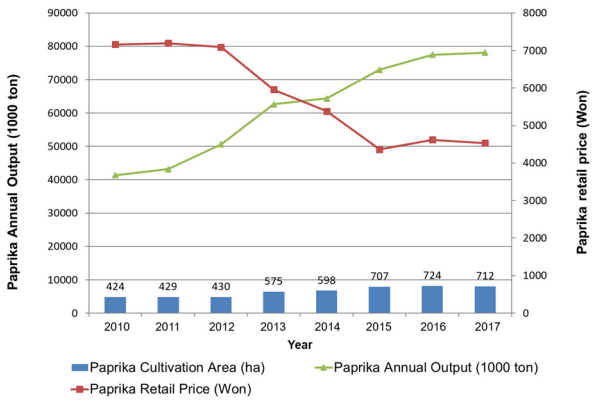

Fig. 1. Trends of paprika cultivation area, annual output, and retail price from year 2010 to 2017

2. METHODS

2.1 Relations between paprika consumption and unstructured big data

2.1.1 Search of paprika-related keywords

In order to search keywords that are related with paprika consumption, multiple sources such as Socialmetrics and Google Trend that were used in one of our previous reports [1]. Paprika-related words were searched by using Socialmetrics solution of Daumsofts (http://www.some.co.kr/) which is one of the commercial big data analysis and visualization platforms, and food/cooking related keywords such as 'sauce' and 'pepper' as well as the name of one of the Korean idols who ate paprika for her weight loss diet were one of the most frequently mentioned related keywords. Paprika-related keywords were also searched by using Google Trends(https://trends.google.com/), and 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, five paprika-related keywords were selected including 'paprika' 'paprika & efficacy,' 'paprika & food,' 'paprika & diet (for weight loss),' and 'paprika & health.' Search frequencies of the five paprika-related keywords were compared over the five years by using Google Trends. The following four paprika-related keywords ('paprika,' 'paprika & efficacy,' 'paprika & food,' and 'paprika & diet,') were selected for further analysis whereas 'paprika & health' was dropped because it has non-noticeable search frequencies as seen in Figure 2.

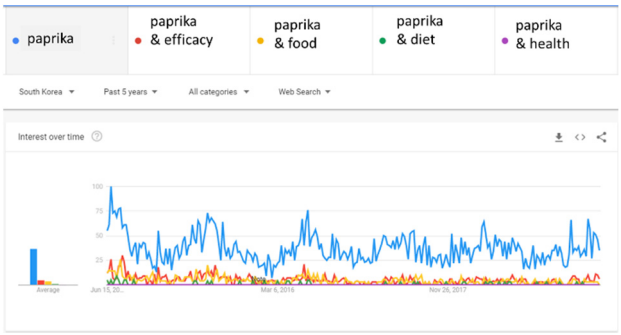

Fig. 2. Search trends of paprika-related keywords for the last 5 years in Google Trend (https://trends.google.com)

2.1.2 Relations between paprika consumption and news, broadcasting programs and SNS

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 blogs in which words "paprika," "paprika and efficacy," "paprika and food," and "paprika and diet" were mentioned between year 2010 and year 2017 as seen in Table 1. Collected data were managed in MongoDB as previously described [2]. The collected data were transformed into weekly format in order to analyze lagged correlation between paprika consumption (amount of money spent to purchase paprika) and keyword frequencies in broadcasting data (Television programs/shows) and social network services data. Lagged correlation between paprika consumption and keyword frequencies in unstructured big data by using ccf function in R.

2.2. Prediction models of paprika consumption

2.2.1 Data used for prediction models of paprika consumption

Agri-food consumers panel data of Rural Development Administration (RDA), wholesale market data of Outlook and Agricultural Statistics Information System (OASIS) of Korea Rural Economic Institute, retail price data of Korea Agricultural Marketing Information Service (KAMIS), pork production data of Korean Statistical Information System (KOSIS) were used as structured data whereas data from

broadcasting programs and blogs were used as unstructured data [1], [2].

Structured data of production and sales of paprika from year 2010 to 2017 were extracted from Mongo database and prepared for analysis in order to predict the demand paprika such as amounts to purchase paprika, daily retail prices of paprika, daily wholesale prices of paprika and so on as seen in Table 2. Agri-food consumers panel data were from Korean consumer panels and have feature of amount to purchase from year 2010 to 2017.

Table 1. Unstructured big data for relations between paprika consumption and news, broadcasting programs and SNS

Cate-gory	Key-word	Arti-cle	Com-ment	Posi-tive term	Negative term	
Broad- cast news	Paprika	1866	2231	7907	384	
	Paprika and efficacy	43	5	330	8	
	Paprika and cooking	208	204	2429	54	
	Paprika and diet	130	102	1046	28	
	Key-word	Pro-gram	View rate	Posi-tive term	Negative term	
Tele-vision pro-grams/ shows	Paprika	411	411	34930	1265	
	Paprika and efficacy	63	63	7891	329	
	Paprika and cooking	354	354	30905	1092	
	Paprika and diet	114	114	12815	542	
	Key-word	Blog	Com-ment	Like	Posi-tive term	Nega-tive term
Blog	Paprika	78813	35354	67877	150440	32749
	Paprika and efficacy	3075	7478	16867	78581	1393
	Paprika and cooking	28320	55505	75484	86487	14970
	Paprika and diet	8408	18758	40300	91384	3969

Table 2. Structured big data for prediction models of paprika consumption

Data name	Feature name	Description
Agri-food consumers panel data	Panel_purchase_amount	Weekly average amount to purchase paprika per consumer panel
Sales of paprika	Retail_price	Weekly retail prices of paprika
	Wholesale_price	Weekly wholesale prices of paprika
	Wholesale_amount_kg	Weekly amounts to wholesale market (kg)
Production of paprika	Output_kg_per_ha	Yield per unit area (ha) per year
	Area_ha	Cultivation area per year (ha)
	Output_kg_year	Output per year (ton)
	Output_kg_year_before	Output in previous year (ton)

Table 3. Unstructured big data for prediction models of paprika

Cate-gory	Feature name	Description	Data No.
Broad-cast news	News_freq	Weekly frequency of keyword mentioned in broadcast news	1,866
	Emotions_Number_Angries	Weekly frequency of comments with angry emoticon of broadcast news where keyword was mentioned	57
	Emotions_Number_likes	Weekly frequency of comments with like emoticon of broadcast news where keyword was mentioned	600
	Emotions_Number_sads	Weekly frequency of comments with sad emoticon of broadcast news where keyword was mentioned	6
	Emotions_Number_wants	Weekly frequency of comments with want more reports emoticon of broadcast news where keyword was mentioned	21
	Emotions_Number_warms	Weekly frequency of comments with moved emoticon of broadcast news where keyword was mentioned	10
	News_comment_freq	Weekly frequency of comment of broadcast news where keyword was mentioned	2,231
	News_positive_term_freq	Weekly frequency of positive term of broadcast news where keyword was mentioned	7,907
	News_negative_term_freq	Weekly frequency of negative term of broadcast news where keyword was mentioned	384
Tele-vision pro-grams / shows	Video_freq	Weekly frequency that keyword was mentioned in television programs/shows other than broadcast news	411
	Video_total_ranking_ave_p	Average television view rate of television programs/shows where keyword was mentioned	411
	Video_freq_times_viewrate	Video_freq times Video_total_ranking_ave_p	411
	Video_positive_term_freq	Weekly frequency of positive term of television programs/shows where keyword was mentioned	34,930
	Video_negative_term_freq	Weekly frequency of negative term of television programs/shows where keyword was mentioned	1,265
Blogs	Blog_freq	Weekly frequency that keyword was mentioned in blog	109,793
	Blog_comments	Weekly frequency of comment of blog where keyword was mentioned	203,025
	Blog_likes	Weekly frequency of comments with like emoticon of blog where keyword was mentioned	335,456
	Blog_positive_term_freq	Weekly frequency of positive term of blog where keyword was mentioned	2,160,130
	Blog_negative_term_freq	Weekly frequency of negative term of blog where keyword was mentioned	45,429

Unstructured data that matches keyword search were collected from broadcasting programs and blogs. Speech from broadcasting programs were converted into text or transcripts were collected. Unstructured data that indicates broadcasting programs and social network services where keywords were mentioned in broadcast news, television programs/shows, and blogs in Korea as seen in Table 3.

Prediction models were developed in order to forecast daily average amount to purchase paprika in year 2017 by

using data from year 2010 to 2016 as training data set and data from year 2017 as a test data set. Structured and unstructured data were used for training and test. Two different algorithms were used to develop prediction models including autoregressive exogenous model and vector error correction model as time-series algorithms.

Autoregressive exogenous (ARX) model is an autoregressive model with exogenous variables and is one of the representative and quantitative dynamics modeling approaches that have been often used in the time series analysis [13], which showed better prediction results in our preliminary study (data not published). In order to predict weekly paprika consumption (amount of money spent to purchase paprika) in year 2017, structured and unstructured data listed in Tables 2 and 3 from year 2010 to year 2016 and arx function (lag selection of 1) in gets package in R were used after daily amounts to purchase paprika were averaged for weekly basis.

Vector error correction model (VECM) developed by Engle and Granger was aimed to have the insertion of short-term adjustments due to the presence of integration [14, 15]. In VECM as well as VAR (Vector Autoregressive) models, more than one variable can be predicted because the interrelations between variables with each other can be seen [13]. In order to forecast weekly paprika consumption (amount of money spent to purchase paprika) in year 2017 by using VECM model and the same dataset as one for ARX, daily amounts to purchase paprika were averaged for weekly basis followed by the granger causality, and cointegration degree test, and lag selection of 1 by using Eviews 10 software.

Predicted paprika consumptions (amounts of money spent to purchase paprika) in year 2017 were compared with actual amounts of money spent to purchase paprika in 2017 in terms of mean absolute error (MAE) and mean absolute percentage error (MAPE) in order to compare accuracy of prediction models given by: $MAE = \frac{1}{n}\sum|y - \hat{y}|$ $MAPE = \frac{100\%}{n}\sum\frac{|y-\hat{y}|}{y}$

3. EXPERIMENTAL RESULTS

3.1 Relations between paprika consumption and unstructured big data

When lagged correlation between purchase amounts of paprika and keywords frequencies from television programs/shows from year 2010 and 2017 was analyzed, there was a statistically significant correlation between positive term frequency in television programs/shows that mentions paprika and diet, and the increase in the purchase amount of paprika with 21~23 weeks of lagged time (maximum at 23 week) as seen in Fig. 3. An example that mentioned paprika and diet in television programs/shows is as follows: "... If you're on a diet, you should eat paprika seeds together… There is capsaicin in the seeds. Capsaicin dissolves body fat. If you are on a diet, please be sure to eat paprika seeds as well…" from a cooking show aired in 2017 from Educational Broadcasting System.

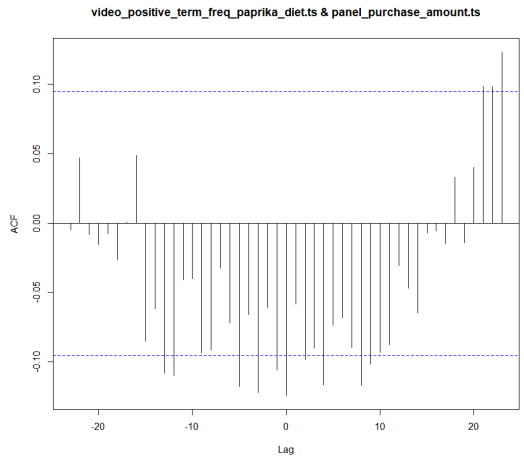

Fig. 3. Correlation between television programs/shows and paprika consumption

When lagged correlation between purchase amounts of paprika and keywords frequencies from blogs from year 2010 and 2017 was analyzed, there is a statistically significant correlation between comment frequency in blogs that mentions paprika and diet, and the increase in the purchase amount of paprika with 15~18 weeks of lagged time (maximum at 18 week) as seen in Fig. 4. Examples that mentioned paprika and diet in blog and comments include "... When I was hungry, I ate cucumbers, paprika and melon a bit… Foods that are helpful when dieting include paprika, bananas, cucumbers and so on..." from a blog and "Have a healthy diet and succeed in diet! I will wait until you upload a picture to prove", "It's a healthy food. Thank you for the info. I'll capture it!" from its comments. However, when lagged correlation between purchase amounts of paprika and keywords frequencies from broadcast news from year 2010 and 2017 was analyzed, statistically significant correlation was not found.

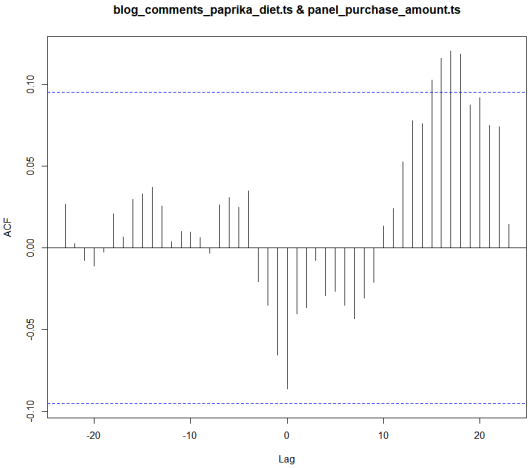

Fig. 4. Correlation between blog and paprika consumption

3.2 Prediction models of paprika consumption

Two time-series algorithms were developed to predict weekly amounts to purchase paprika in year 2017, and model accuracy was compared by using MAPE and MAE. Two

different data sets were used to develop prediction models, which include structured and unstructured data and structured and unstructured data as well as paprika and diet related unstructured data shown in Table 1 as seen in Table 4. Between the two time-series models with structured and unstructured data, ARX model showed better prediction with lower MAPE and MAE. When paprika and diet related unstructured data were added to the dataset, the MAPE and MAE of ARX model became lower, which improved the prediction accuracy.

Table 4. Prediction models of weekly paprika consumption by using the amounts to purchase paprika in year 2017

Model	Variables	Mean Absolute Percentage Error (%)	Mean Absolute Error
Vector Error Correction model	Structured and unstructured data	14.72	376.14
Auto-regressive exogenous model		10	257.41
Auto-regressive exogenous model	Structured and unstructured data as well as paprika and diet related unstructured data shown in Table 1	9.89	250.93

Two ARX models, one with structured and unstructured data and the other with structured and unstructured data as well as paprika and diet related unstructured data, compared with actual amounts in graph as seen in Figure 5. The patterns of ARX with the former dataset and ARX with the latter dataset stay close to each other. However, when the actual weekly amounts dropped sharply at week 37(where the arrow indicates in Figure 5), ARX with the latter dataset mimicked the pattern of actual weekly amounts whereas ARX with the former dataset did not.

Figure 5. Comparison of predicted weekly paprika consumption by using the amounts to purchase paprika in year 2017

4. CONCLUSION

In this paper, we analyzed relations of SNS and broadcasting program on paprika consumption in order to identify potential factors that can promote paprika consumption in advance. We identified that there are statistically significant correlations between paprika consumption in the aspect of amount to purchase paprika and television programs/shows and blogs that mentioned paprika and diet with lagged times whereas no significant correlation found between purchase consumption and broadcast news. Based on the findings, television programs/shows and blogs could be targeted to promote paprika consumption. When paprika consumption promoting is made through broadcasting programs and blogs, it would be better to highlight paprika for diet.

Our findings suggested that he lagged times between purchase consumption in the aspect of amount to purchase paprika and television programs/shows and blogs that mentioned paprika and diet as 21~23 weeks and 15~18 weeks, respectively. The suggested lagged times were considered long compared to the onion consumption (5 weeks for purchase and blog, 11 weeks for purchase and television programs/shows) in our previous reports and a paper on internet search indexes and agri-food products purchases (5 to 6 days in the 30 agri-food products tested) [1], [16]. Possible speculations may include that paprika is not considered as an essential agricultural product in Korea and purchasing paprika for weight loss diet can take long time from plan to action, which are to be proven in further study.

From the prediction models development of paprika consumption by using structured and unstructured big data, we identified that adding paprika and diet related unstructured data to the prediction model improved the model predictability in the ARX model tested.

Some of the limitations may include that the correlations between paprika consumption and television programs/shows and blogs that mentioned paprika and diet with lagged times are to be verified in a cause-and-effect relationship so that the claimed correlations are not coincidental and the claimed correlations could be used for promotion when paprika are overproduced.

In this paper, we identified that statistically significant correlations between paprika consumption and television programs/shows and blogs that mentioned paprika and diet with lagged times, which could be used for paprika consumption promotion. We also identified that adding paprika and diet related unstructured data to the prediction model improved the model predictability in the prediction model tested. Although there are a few papers that used structure and unstructured data for prediction in agricultural field, this is the first report to predict consumption of agricultural products by using structured and unstructured data as far as the authors are aware [1], [11], [16].

ACKNOWLEDGEMENT

This research was funded by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA), grant number 319003-01.

REFERENCES

[1] H. Rah, E. Oh, D. I. Yoo, W. S. Cho, A. Nasridinov, S. Park, Y. Cho, and K. H. Yoo, "Prediction of Onion Purchase Using Structured and Unstructured Big Data," The Journal of the Korea Contents Association, vol. 18, 2018, pp. 30-37. doi: 10.5392/JKCA.2018.18.11.030

[2] H. Rah, K. Park, B. An, S. Choi, D. Chae, and K. H. Yoo, "Development of Prediction Model of Agro-Food Demand by Unstructured and Structured Bigdata," The 5th International Conference on Big Data Applications and Services Proceeding, 2017, pp. 122-127.

[3] M. H. Shin, S. H. Oh, D. Y. Hwang, S. S. Seo, and Y. C. Kim, "Effect of SNS Characteristics on Consumer Satisfaction and Purchase Intention of Agri-food Contents," JOURNAL OF THE KOREA CONTENTS ASSOCIATION, vol. 12, 2012, pp. 358-367. doi: 10.5392/JKCA.2012.12.11.358

[4] S. H. Kim, *The Impact of Foot-and-Mouth Disease on Pork Consumption: Analysis of Consumer Response to Media*, Master Thesis, The Graduate School, Seoul National University, 2016.

[5] C. Artola, F. Pinto, and P. de Pedraza García, "Can internet searches forecast tourism inflows?," International Journal of Manpower, vol. 36, 2015, pp. 103-116. doi: 10.1108/IJM-12-2014-0259

[6] H. Choi and H. Varian, "Predicting the present with Google Trends," Economic Record, vol. 88, 2012, pp. 2-9. doi: 10.1111/j.1475-4932.2012.00809.x

[7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," Journal of Computational Science, vol. 2, 2010, pp. 1-8. doi: 10.1016/j.jocs.2010.12.007

[8] K. Kurumatani, "Time Series Prediction of Agricultural Products Price Based on Time Alignment of Recurrent Neural Networks," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 81-88. doi : 10.1109/ICMLA.2018.00020

[9] J. Kim, M. Cha, and J. G. Lee, "A Model for Nowcasting Commodity Price based on Social Media Data," Journal of KIISE, vol. 44, 2017, pp. 1258-1268. doi: 10.5626/JOK.2017.44.12.1258

[10] X. V. Meza and H. W. Park, "Organic Products in Mexico and South Korea on Twitter," Journal of Business Ethics, vol. 135, 2016, pp. 587-603. doi: 10.1007/s10551-014-2345-y

[11] D. I. Yoo, "Vegetable Price Prediction Using Atypical Web-Search Data," in 2016 Annual Meeting, July 31-August 2, 2016, Boston, Massachusetts, 2016. doi: 10.22004/ag.econ.236211

[12] H. Rah, E. Oh, W. S. Cho, K. H. Yoo, and Y. Cho, "Paprika Purchase Prediction By Using Structured and Unstructured Big Data," The 7th International Conference on Big Data Applications and Services Proceeding, 2019, pp. 238-242.

[13] K. Fukata, T. Washio, K. Yada, and H. Motoda, "A method to search ARX model orders and its application to sales dynamics analysis," in *Data Mining for Design and Marketing*, Chapman and Hall/CRC, 2009, pp. 90-103.

[14] A. Suharsono, A. Aziza, and W. Pramesti, "Comparison of vector autoregressive (VAR) and vector error correction models (VECM) for index of ASEAN stock price," in AIP Conference Proceedings, 2017, p. 020032. doi: 10.1063/1.5016666

[15] J. Zhang, W. Hu, and X. Zhang, "The Relative Performance of VAR and VECM Model," in 2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, 2010, pp. 132-135. doi: 10.1109/ICIII.2010.195

[16] H. Rho, S. Y. Kim, and T. Y. Kim, "Does the Internet Search Index Precede Agri-Food Product Purchases?," Journal of Rural Development, vol. 42, 2019, pp. 1-34.

Yongbeen Cho
He received his MA degree in agricultural education from Seoul National University in 2002. He joined Rural Development Administration, Republic of Korea where he works as a senior researcher. His research interests include application of agricultural big data for agricultural policy.

Eunhwa Oh
She received her BS degree in computer science from Hannam University, Korea in 2018 and joined Department of Big Data in Chungbuk National University to pursue her MS degree. Her research interests include database.

Wan-Sup Cho
He received BS degree from Kyungpook National University, Korea in 1985, MS and PhD in Computer Science from Korea Advanced Institute of Science and Technology in 1987 and 1996, respectively. He is currently a professor in Department of Management Information Systems, Chungbuk National University. His research interests include database, business intelligence, enterprise resource planning, and big data.

Aziz Nasridinov
He received BS degree in information technologies from Tashkent University of Information Technologies, Uzbekistan, in 2006, and MS and PhD degrees in computer engineering from Dongguk University, South Korea. He is currently an associate professor in Data Analytics laboratory, Department of Computer Science, Chungbuk National University. In the past, he has worked as a post-doctoral researcher at Sookmyung Women's University, and as a research professor at Dongguk University. His research interests include database systems, data mining and parallel and

distributed computing. He has published more than 10 scientific papers in various international journals. He is also an editorial board member of several international journals.

Kwan-Hee Yoo

He received BS degree in Computer Science from Chonbuk National University, Korea in 1985, MS and PhD in Computer Science from Korea Advanced Institute of Science and Technology in 1988 and 1995, respectively. He is currently a professor in Department of Computer Science, Chungbuk National University. His research interests include u-Learning system, computer graphics, 3D character animation, dental/medical applications.

HyungChul Rah

He received BS degree in Veterinary Medicine from KonKuk University, Korea in 1996, and BS in Forest Management from Korea University, Korea in 2000, MPVM in Preventive Veterinary Medicine and PhD in Comparative Pathology from Univ. of California, Davis, USA in 2001 and 2006, respectively. He is currently a visiting professor in Department of Big Data Convergence. His research interests include big data analysis in agriculture and healthcare.