Medical Advertising Content Filtering System in Online Knowledge Sharing Service

Yoosin Kim[1], Tae Yun Kim[2], Sang Hyun Choi[2]

[1] Big Data Analytics Dept., University of Seoul, Korea
[2] Department of MIS, BK21+Team, Chungbuk National University, Korea
yoosin25@uos.ac.kr, kim-taeyoun@hanmail.net, chois@cbnu.ac.kr

Abstract

Online knowledge sharing service such as Q&A communities in the Web is a representative service as collective intelligence among people and a critical web-service to entice online consumers as well. A huge volume of Q&A content is generated and shared in real time, however there are also uncertain information including not only valuable knowledge but also commercial data such as advertisements, marketing, and even wrong information. That commercial content is able to lead online-users to controversy. This study proposes a content filtering system in the knowledge sharing community to classify whether the information is commercial or not. The filtering system applies linguistic feature sets and employs a support vector machines algorithm in machine learning methods to classify whether the information in the knowledge sharing community is commercial or not. To build the algorithm and validate the system, we set the target domain within the healthcare content and gathered question and answer content about lung cancer of knowledge sharing service in a Korean web-portal site, Naver.com. As the result, the proposed system accomplished accuracy average 84.0% with the Ads words and the document length.

Keywords: Lung cancer, Q&A community, Machine learning, Text mining

1 Introduction

Emerging Web 2.0 technology in the internet, communication and sharing information among online users has been increasing more and more. Web service providers such as Yahoo, Naver, and Weibo, regarded the customer's favorite as a business opportunity, and then opened community service as the knowledge sharing site. The knowledge is composed by the broad and informal terms in various domains from daily life to high technology. And also the knowledge is produced by a group of people, from the armature to the professionals who are interested in the question. In other words, online knowledge is an expanded concept including our daily life, common sense and advice, and a relatively unstable and fluid knowledge provided directly by producer of a variety of knowledge, including the user. Those service sites were very popular and a representative service of web service providers and currently the knowledge sharing service became a new mechanism as building and sharing the knowledge and finally seem to be recognized as collective intelligence to solve a difficult problem. Collective intelligence is considered as an innovative and brand-new knowledge creation model [1]. According to Pierre Levy (1994), collective intelligence exists in everywhere, has given value constantly, is adjusted in real time, and is mobilized by practical skills [2].

However, it has been pointed out due to limitations such as precision and reliability of answers [1, 3-5]. First, despite of group decision, the knowledge in collective intelligence could be incorrect. In consensus for solving a problem and making group decision, there are much collaboration and competitions among people. Sometime the works might be much conflict and drive the solution to the wrong way. In addition, the biased information and interpretation could provide the inaccurate knowledge to the participator. Last, there is no such a clear control system for these collaboration works and coordination. For example, several web-service businesses have provide online community service as knowledge sharing sites such as Naver Q&A (Jisikin), Yahoo Answers, and Daum Agora. However they do not control in making group decision as collective intelligence and thus confront the same problem by wrong information and absence of a control system.

In the knowledge sharing service, various subjects are discussed from just interests and tips to professional area such as economy, phycology, politics, and healthcare. Particularly, medical information is much sensitive to people because questions tend to seek effective solutions for the sick and disease. In this regard, medical information in collective intelligence has to be careful to generate and share the knowledge

*Corresponding Author: Sang Hyun Choi; E-mail: chois@cbnu.ac.kr
DOI: 10.3966/160792642018051903024

since the wrong information would drive someone in bad situation. In this study, we interested whether the medical information is correct and proper to the questions, and then tried to propose a system for filtering the inaccurate content such as commercials. The filtering system applies linguistic feature sets and employs a support vector machines algorithm in machine learning methods to classify whether the information in the knowledge sharing community is commercial or not. To build the algorithm and validate the system, we set the target domain within the healthcare content and gathered question and answer content about lung cancer in knowledge sharing service of Korean web-portal site Naver.com.

This study suggests an intelligent filtering system to classify and remove the commercial content seducing users in online knowledge sharing service.

2 Related Works

2.1 Knowledge Sharing Service in the Web

Web 2.0 technologies in the internet have improved communication and sharing information among online users. Online communities and SNS have been popular to people because of the usefulness of group decision as knowledge sharing service. Various subjects are discussed in them from interests and tips to professional area such as economy, phycology, politics, and healthcare. The information in the service is composed by the broad and informal terms in various domains from daily life to high technology. And also the knowledge is produced by a group of people, from the armature to the professionals who are interested in the question. In other words, online knowledge is an expanded concept including our daily life, common sense and advice, and a relatively unstable and fluid knowledge provided directly by producer of a variety of knowledge, including the user.

In that reason, web service providers such as Yahoo, Naver, and Weibo, regarded the customer's favorite as a business opportunity, and then provided and focused the knowledge sharing service as a representative service with Naver Jisik-in, Yahoo Answers, and Daum Agora. Especially, Naver Jisik-in is a critical service of Naver.com, which is market share No.1 in South Korea. Naver, currently taking over 70% market share in web-portal business, had increased market share rapidly increased with Jisik-in service when their market share was stated around 30% early in 2000. Jisik-in is a compound Korean word meaning the intellectual. Jisik-in was opened at October 7 2002 and reached a hundred million questions, 29 million login-users, 13 million questioners and 7 million answers for 10 years at 2012 (Figure 1).

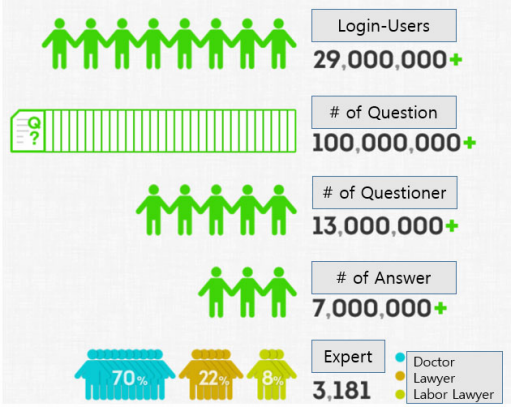


Figure 1. Achievement of Jisik-in for 10 years

The service was growing gradually a new mechanism as building and sharing the knowledge. Nowadays, exploring the knowledge sharing site is very common work to find something to know and figure out solutions of problems and was recognized as collective intelligence that solves a difficult problem. As the result, collective intelligence became an innovative and brand-new knowledge creation model [1, 6].

By the way, the users concern to use the information in the knowledge sharing service as well. User's interests in online communities and SNS are influenced by the forwarding and comment behaviors [17]. Therefore incompleteness and limitation such as precision and reliability of answers is rising as larger as their usefulness and popularity [1, 3-5].

First, the knowledge in collective intelligence might be incorrect. Secondly, the wrong information might be misunderstood as the correct knowledge. Third, there is no such a clear control system for these problems. For instance, the knowledge sharing service deals with various subjects are discussed from just interests and tips to professional area such as economy, phycology, politics, and healthcare. Particularly, medical information is much sensitive to people because questions tend to seek effective solutions for the sick and disease. In this regard, medical information in collective intelligence has to be careful to generate and share the knowledge since the wrong information would drive someone in bad situation. In this study, we interested whether the medical information is correct and proper to the questions, and then tried to propose a system for filtering the inaccurate content such as commercials.

For example, knowledge sharing sites such as Naver Jisik-in, Yahoo Answers, and Daum Agora, are serviced through collective knowledge, and they also confront the same problem by wrong information and absence of a control system. Particularly, the knowledge asking more sensitive and accurate information such as medical information, has to be careful to generate and share the knowledge [5, 7]. In this regard, much research has suggested the methods to identify credibility and reliability of question-answer content in online knowledge sharing service [1, 5-6, 8].

2.2 Korean Natural Language Processing

In this study, we target Naver Jisik-in, which is the No.1 the knowledge sharing service in South Korea. The knowledge content is written by Korean. Korean language has much complicate structure to analyze corpora and syllables from text [9]. To deal with the content, Korean natural language processing (KNLP) is necessary. There are many studies using KNLP, and then they commonly conducted a few methods such as eliminating garbage, extracting features, parsing letters, and tagging with characters [10-12]. In natural language content, there are many kinds of useless characters such as emoticons, numbers, punctuations and stop words, and they would be barriers to extract information from contents. Thus, the researchers first removed these obstacles from experiment data and eliminated stop-words (e.g., I, me, my, and mine) and meaningless common words for the efficiency of analysis. In fact, we could not consider other kind language such as English, Chines, and so on because NLP is really depending on the type of language.

2.3 Document Classification

To filter the garbage content such as e-mail spam and mobile phone spam message, filtering methods using document classification are frequently used [8]. The methods typically employ filtering algorithms such as Bayesian classifier, logistic regression and decision tree. Meanwhile a mobile spam filtering study proposed a simple and light method just using keyword frequency ratio to save mobile device resources such as storage space, CPU, memory, etc. [8]. Another research attempted to modify naïve Bayes classifier which splits question-answer documents into information, opinion, and suggestion using structural characteristics of the document [13]. Most of the current document classification system is mainly applied to the filtering method based on the message rule and e-mail filtering system applying many other stochastic methods has been developed.

Rule-based classification. The filtering method using the message rule can determine whether spam or not finding words than can represent the characteristics of spam and it is simpler than the stochastic method. Also, it is possible to obtain a relatively good performance even in the recall and precision. On the other hand, the user has to enter a direct message rules and it must be constantly updated as spam mail changes its type. Besides, there is a limit to be an accurate filtering because it has cases of two, only true or false.

Machine learning method based classification. Machine learning is a subfield of artificial intelligence and computer science, and deals with algorithms and techniques that improve automatically through experience rather than follow programmed instructions. Machine learning is applied into various tasks from data mining programs that discover general rules in large data sets, to information filtering systems. There are various machine learning methods, but we employed four methods which are frequently adopted in previous research: NB, DT, NN, and SVM. NB is one of frequently employed classifiers for text mining because of their simplicity. DT learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Artificial neural network method is a learning algorithm that is inspired by the structure and functional aspects of biological neural networks.

3 Commercial Content Filtering System

3.1 System Structure

We propose an intelligent text mining system for filtering commercial content in online healthcare Q&A community. The system is composed of pre-processing, parsing, tagging, and filtering module which followers previous research [11, 14]. A classification algorithm of the filtering module applies SVM methods using linguistic features Figure 2 presents process and functions of the proposed system in detail.

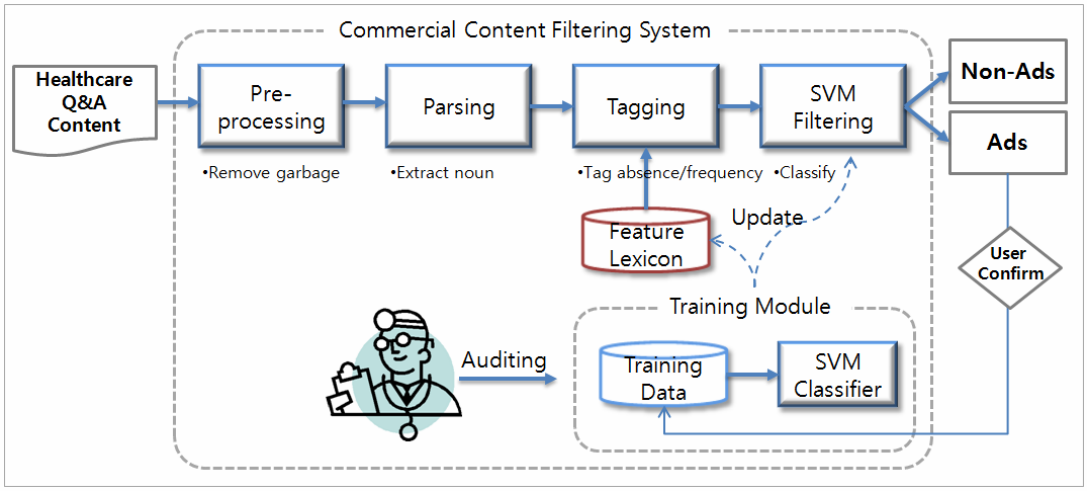


Figure 2. Overview of commercial content filtering system

Pre-processing module. The first step for the filtering system is to make the content clean due to increase the accuracy of commercial data classification. The user generated content in online Q&A community has several types of unnecessary objects such as html tags, punctuation, numbers, and emoticons. In the pre-processing module, those meaningless letters are eliminated by command sources of "tm" and "KoNLP" packages in R program.

Parsing module. Parsing module is in charge of extracting noun from the sentence. The parser splits a sentient up into each word and then selects just nouns in them. In this working, the system employs the command and linguistic source in KoNLP.

Tagging module. In a bag of word after parsing, meaningful nouns are remained for a next step, tagging. Tagging module determines whether a word is useful as a feature for the classification algorithm using a dictionary which consists the feature terms. If an extracted word matched as a feature term, it would be tagged as exist condition (1) or else absence (0).

Filtering module. Filtering module does remove the commercial content by the classification algorithm. In this step, various classification methods such as the rule-based classification, the naïve Bayesian classifier, the neural network algorithm, and the support vector machines classifier can be used for filtering the commercial content. The proposed system employs support vector machines (SVM) classifier, which has driven generally higher performance at text categorization in two category case [15].

Training module. For tagging and filtering module, a defined feature dictionary and a classification algorithm are needed. Train module is in charge of generating and learning the dictionary and filtering algorithm using training data. In addition, training module is learned incrementally for the advanced dictionary and classifier as well.

4 Experiment And Result

4.1 Data

The healthcare data was collected from "Jisik-in (http://kin.naver.com)" the knowledge sharing service in Naver which is the popular search portal website in South Korea. Jisik-in means the much intellect people and the Q&A content in the Jisik-in service has been accumulated over 229 million since 2002 year. The portal web site has captured 70% of the web-portal service area and been ranked as the number one site in Korea over 10 years. Since there is much healthcare information in the online Q&A community, we chose a specific subjective, Lung Cancer, which is watched as a critical decease. We made query with "lung cancer" keyword in Jisik-in and crawled questions and answers written by online users. Through the crawling, we gathered 217 questions and their answer 535 documents from January 1 in 2012 year to May 31 in 2015 year.

We tried to analyze and understand the questions and answers in context, and thus found that the content can be categorized by the context of the healthcare information. Therefore, we first divided the content into two group, commercial content and normal healthcare content. And then, we attempted to categorize the healthcare content into four types: the symptoms of lung-cancer, the treatments, the management, and the prognosis. In this work, we read all document and categorized them by the types in manual way.

As the result of manual labeling, the commercial content shared 32% of the collected data and then the treatments of lung-cancer were mentioned within 9% of the data. Offering words of consolation and private experiences to the questioner was posted frequently in 15% of ratio.

Table 1. Result of classification

	Type	Amount (n=535)	Ratio (%)
Commercial		167	31%
Non-Commercial	Diagnosis	80	15%
	Treatments	73	14%
	management	67	13%
	Advice	118	22%
	The rest	30	6%

4.2 Feature Extraction

Machine learning methods including SVM classifier require training data and features for the classification. In a previous study, researchers applied several kinds of variables such as various sizes of terms, n-grams, frequency, positions, and the experiment achieved a performance of 81.4% by the SVM classifier using the feature dictionary with about 2000 unigram terms [19]. In the proposed system, we defined five features: length of a document, frequency of religion words, frequency of food words, frequency of therapy words, and frequency of Ads words.

First, the document length means a total number of characters in an answer document. In general, the true answer from online user tends to be simple and short, but much commercial content from marketers and business users shows long sentence including detailed explanations such as causes of disease, authorities on cancer, famous hospitals, and so on. Next, religion

words and food words are a kind of stop words. Regarding the healthcare Q&A as the collective intelligence in the specialized domain, a certain type of words is not proper for the medical information and knowledge. The religion words asking to believe the God and the food words as the healthy material are not related to the healthcare information. On the other hand, the therapy terms of medical domain can be helpful to recognize whether a document is linked with healthcare information. For making the lexicon of religion, food, and therapy words, we run the parser to the collected data and selected the religion, food, and therapy words from the extracted nouns in manual way. Last, Ads words feature set is the group of terms only existing in commercial content. For generating them, we parsed commercial labeled content and extracted 3,000 high frequency words and did same way with non-commercial labeled content. After comparing with two groups, we remained the Ads words only existing in the commercial part. Table 2 is a sample of religion, food, therapy, and Ads words.

Table 2. Sample words in feature dictionaries

Dictionaries	Sample words	# of Feature
Religion	God, Bible, Jesus, Buddha	62
Food	Healthy food, Fermented food, Chaga mushroom, Cabbage, Mineral water	109
Therapy	Radiotherapy, Adenosine, AmyNex, BBRC, Calebin	293
Ads Words	Alternative medicine, Fork remedy, Expectant treatment, Fear	1436

4.3 SVM Classification Result

Our goal in SVM filtering work is to classify whether a user-generated answer in knowledge sharing service is commercial content with purpose to entice the healthcare customer. For this machine learning method, we made a sample data merged with 150 non-commercial answers and 150 commercial data. We calculated features of each answer document in the sample data using the above feature dictionaries and then conducted classification experiments by SVM algorithm. In addition, we applied a k-fold cross-validation estimator in to the SVM experiment in order to less over-fitting problems in algorithm learning. The k-fold cross-validation provides a lower variance than the holdout method which is consisted a training data set and a validation (or test) data set. The k-fold cross-validation first divides an original data set into k-equal size subsamples and then repeats the holdout method in k-times. Each time, one subsample of the k-subsamples becomes the validation data set and the rest k-1 subsamples put the training data set together. Through this repeated works, the k-fold validation estimator can reduce the variance in different partitions of the data to form training and test sets by averaging over K different partitions, thereby performance estimate is less sensitive to the partitioning of the data.

The performance of algorithms and feature sets can be evaluated by several statistical measures such as accuracy, recall, precision, and F-measure [9, 16-18]. In this study, we used the total accuracy which is defined as the percentage of sentiments correctly predicted of total instances:

$$Accuracy\ \alpha = \frac{instances\ correctly\ predicted}{Total\ Instances}$$

According to the result of 10 fold cross-validation experiment (see Table 3), we can see the 3rd experiment with the Ads words and the document length achieved the best performance of accuracy average (84.0%) in every test cases. Interestingly, a case using only the Ads words stayed around 76% of accuracy average and also another test with the document length did not reach to 60% of accuracy average. However the additional experiment merging two features, the Ads words and the document length, obtained remarkable performance against individual accuracy of each feature, thereby we can contemplate the synergy effect of both features. Particularly, the document length feature showed low-accuracy in the single use but high-performance in the combining test with other features. Therefore the filtering system needs to consider involving the document length feature in the classification algorithm. On the other hand, the religion words, the food words, and the therapy words stayed less performance than the case of merging the Ads words and the document length (84.0%), despite combining with the Ads words. In a test with three features, the therapy words, the Ads words, and the document length, the accuracy (83.3%) is not higher than the result (84.0%) merging the Ads words and the document length. The result intends that the Ads words can work efficiently without other textual features, and the system has to make effort to build the optimized commercial words dictionary in the target domain as well.

Table 3. SVM 10-fold cross-validation result

Ex	Features					Accuracy (%)		
	Ads	Length	Religion	Food	Therapy	Min	Max	Average
1	V					66.7	86.7	76.0
2		V				50.0	70.0	59.3
3	V	V				80.0	90.0	84.0
4	V		V	V	V	66.7	86.7	76.0
5	V	V	V	V	V	76.7	86.7	81.0
6	V		V			76.7	86.7	81.3
7	V			V		73.3	86.7	81.0
8	V				V	76.7	90.0	82.3
9	V	V			V	80.0	86.7	83.3

In this experiment, we used "tm" and "KoNLP" packages in R project to calculate the feature score and also applied the SVM package "e1091" to learn the classification algorithm.

5 Conclusion

After emergency of the knowledge sharing service as collective intelligence in the Web, online Q&A communities such as Yahoo Answer and Naver Jisik-in became a very popular service among people. It also takes a roll as a critical service in the web-portal site to entice online consumers. A huge volume of question and answer content is generated and shared in real time. Nowadays, the knowledge sharing service collective intelligence is regarding as an innovative and brand-new knowledge creation model.

However, its usefulness and popularity is growing rapidly, concerns about limitations such as precision and reliability of answers is rising as well. Online users indeed can see the uncertain information including not only valuable knowledge but also commercial data such as advertisements, marketing, and even wrong information. Since that commercial content is able to lead online-users to controversy, the garbage data filtering system is needed to prohibit misleading the customer about the healthcare information.

In this paper, we aim to (1) propose the intelligent system to pick out the commercial content in online Q&A community, and (2) build and validate text mining classification algorithm of the filtering system. The commercial content filtering system is composed of pre-processing of document, parsing text and extracting nouns, tagging features, filtering commercial data, and algorithm training module. A classification algorithm of the filtering module applies SVM methods using linguistic features such as: character length of a document, frequency of religion words, frequency of food words, frequency of therapy words, and frequency of Ads words.

We collected the healthcare data mentioning "Lung cancer" in the online Q&A community of Naver, a market leader in Korean portal websites. The data composed 217 questions and their answer 535 documents from January 1 in 2012 year to May 31 in 2015 year. In the collected data, we sampled 300 answer documents with 150 non-commercial and 150 commercial data for the validation experiment of the system and conducted classification experiments by SVM algorithm.

We demonstrated various combinations with features, and the case using just two features, the Ads words and the document length, achieved the highest performance with 84.0% of accuracy average in the 10 fold cross-validation. Through the experiments, we found the document length feature did not perform a high accuracy by itself, but showed better performance in combining with other features. Therefore the filtering system needs to consider involving the document length feature in the classification algorithm. On the other hand, the religion words, the food words, and the therapy words did not reach the use of the Ads words feature. Consequently, the Ads words and the document length feature efficiently worked in classifying the wrong healthcare information, and the optimized commercial words dictionary in the target domain is much important to increase the filtering performance in the system.

We expect that our proposed system provides several implications and contributions. First of all, the proposed system provides practical advantages to web service providers and the users. The service providers can obtain a control tool to improve their service quality by the proposed system that filters the wrong healthcare information in online knowledge sharing service. And also the knowledge seekers can take more accurate and useful information without garbage. Second, our investigation with various features showed which feature works efficiently to increase the classification performance. In addition, the experiment result implies that the researchers should consider which linguistic feature set and how many terms in the feature are needed in the filtering system. Last, we used open source packages of the R project for demonstrating Korean NLP and machine learning methods, thus potential users can consider these tools and techniques for immediate adoption. We believe this article can support practical and reliable reference to commercial content filtering in healthcare domain as well as other domain knowledge sharing service.

This study also has several challenges to be improved. Even though there are many kinds of healthcare information, we just tested lung cancer in Q&A community because of the limited knowledge in the medical field. It means our result should lean toward the general knowledge and thus this research needs to be validated by the expert such as the medical doctor. Another challenge is to gather huge volume of healthcare knowledge data as real Big-data and employ various features such as user population rank, number of hits, and the author. Last, future research needs to expand to evaluate the quality of each answer document in online Q&A community. It will be more helpful the online users to make the choice of the information. In addition, even though collective intelligence is built by various language such English, Chines, and Japanese, we just treated Korean knowledge because of limitations of natural language processing technology. Therefore, future studies can analyze various language of content and compare among them.

Acknowledgments

Following are results of a study on the "Leaders Industry-university Cooperation" Project, supported by the Ministry of Education, Science & Technology (MEST). This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2016-H8501-16-1013) supervised by the IITP(Institute for Information & communication Technology Promotion). This work was supported by the intramural research grant of Chungbuk National University in 2015 as well.

A preliminary version of this paper was presented at ICONI 2015, and was selected as an outstanding paper.

References

[1] D. Kim, G. Park, S. Lee, QualityRank: Measuring Authority of Answer in Q & A Community Using Social Network Analysis, *Journal of KISS: Databases*, Vol. 37, No. 6, pp. 343-350, December, 2010.

[2] P. Fichman, A Comparative Assessment of Answer Quality on Four Question Answering Sites, *Journal of Information Science*, Vol. 37, No. 5, pp. 476-486, October, 2011.

[3] Z. Shi, H. Rui, A. Whinston, Content Sharing in a Social Broadcasting Environment: Evidence from Twitter, *MIS Quarterly*, Vol. 38, No. 1, pp. 123-142, March, 2014.

[4] S.-J. Kim, Research Trends of the Credibility of Information in Social Q&A, *Journal of the Korean Society for information Management*, Vol. 29, No. 2, pp. 135-154, June, 2012.

[5] S.-J. Kim, Answerers' Strategies to Provide Credible Information in Question Answering Community, *Journal of the Korean Society for information Management*, Vol. 27, No. 2, pp. 21-35, June, 2010.

[6] H. Yu, Y. J. Seo, H. R. Ji, An Investigation of the Communication Style Done by Patients and Medical Professionals in Jisikin at www.naver.com, *Health Communication Research*, Vol. 11, pp. 49-73, December, 2014.

[7] S.-E. Kim, J.-T. Jo, S.-H. Choi, SMS Spam Filtering Using Keyword Frequency Ratio, *International Journal of Security and Its Applications*, Vol. 9, No. 1, pp. 329-336, January, 2015.

[8] J.-S. Lim, J.-M. Kim, An Empirical Comparison of Machine Learning Models for Classifying Emotions in Korean Twitter, *Journal of Korea Multimedia Society*, Vol. 17, No. 2, pp. 232-239, February, 2014.

[9] Y. Kim, D. Y. Kwon, S. R. Jeong, Comparing Machine Learning Classifiers for Movie WOM Opinion Mining, *KSII Transactions on Internet & Information Systems*, Vol. 9, No. 8, pp. 3169-3181, August, 2015.

[10] Y. Kim, S. R. Jeong, Opinion-Mining Methodology for Social Media Analytics, *KSII Transactions on Internet and Information Systems*, Vol. 9, No. 1, pp. 391-406, January, 2015.

[11] I. Ghani, S. R. Jeong, A ROle-Oriented Filtering (ROOF) Approach for Collaborative Recommendation, *Enterprise Information Systems*, Vol. 10, No. 7, pp. 697-728, September, 2016.

[12] J. Yeon, J. Shim, S.-G. Lee, Modified Naïve Bayes Classifier for Categorizing Questions in Question-Answering Community, *Journal of KIISE: Computing Practices and Letters*, Vol. 16, No. 1, pp. 95-99, 2010.

[13] Y. Kim, R. Dwivedi, J. Zhang, S. R. Jeong, Competitive Intelligence in Social Media Twitter: iPhone 6 vs. Galaxy S5, *Online Information Review*, Vol. 40, No. 1, pp. 42-61, February, 2016.

[14] Y. Kim, S. R. Jeong, I. Ghani, Text Opinion Mining to Analyze News for Stock Market Prediction, *International Journal of Advances in Soft Computing and its Application*, Vol. 6, No. 1, pp. 1-13, March, 2014.

[15] M.-A. Mittermayer, G. F. Knolmayer, NewsCATS: A News Categorization and Trading System, *Sixth IEEE International Conference on Data Mining*, Hong Kong, China, 2006, pp. 1002-1007.

[16] H. Rui, Y. Liu, A. Whinston, Whose and What Chatter Matters? The Effect of Tweets on Movie Sales, *Decision Support Systems*, Vol. 55, No. 4, pp. 863-870, November, 2013.

[17] Y. Liu, W.-G. Yuan, User Posting Behavior Analysis and Modeling in Microblog, Journal of Internet Technology, Vol. 16, No. 5 , pp. 811-815, September, 2015.

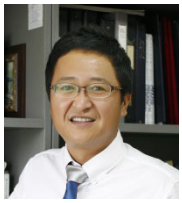
[18] P. Levy, *L'intelligence collective*, Éditions La Découverte, 1994.

[19] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment Classification Using Machine Learning Techniques, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, PA, 2002, pp. 79-86.

Biographies

Yoosin Kim is a Big-data Analytics Visiting Professor in the University of Seoul. He received a Ph.D. with a research for Stock Index Prediction of News Big-data from Kookmin University in Seoul, Korea. He was a post-doctoral researcher in the University of Texas at Arlington, a data scientist at Accenture, and business analyst at SK. He has studied a fire-risk prediction model, Social Economic Index, a public service process mining methodology, and big data analytics.

Tae Yun Kim is a Data scientist in the Funnywork Corporation of Seoul. He received a Master degree in management information systems from Chungbuk National University in Cheongju, Korea. He has consulted for a number of organizations including Ministry of the interior and Statistics Korea. His current research topics include process mining, visualizations and infographics, and big data analytics.



Sang Hyun Choi is a professor at the Department of Management Information Systems in the Chungbuk National University. He received the Ph.D. degree in management information systems from Korea Advanced Institute of Science and Technology in Seoul, Korea. From 1998 to 2002, he was a Senior Consultant at the Entrue Consulting, LG CNS. His research interests include big data analytics, recommendation systems, data mining, and information systems planning.