Intelligent Automation And Soft Computing, 2020
Copyright © 2020, TSI® Press
Vol.26,no.1,141–147
https://doi.org/10.31209/2019.100000135

Word Embedding Based Knowledge Representation with Extracting Relationship Between Scientific Terminologies

Mucheol Kim, Junho Kim, Mincheol Shin

School of Computer Science and Engineering, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul, Korea

ABSTRACT
With the trends of big data era, many people want to acquire the reliable and refined information from web environments. However, it is difficult to find appropriate information because the volume and complexity of web information is increasing rapidly. So many researchers are focused on text mining and personalized recommendation for extracting users' interests. The proposed approach extracted semantic relationship between scientific terminologies with word embedding approach. We aggregated science data in BT for supporting users' wellness. In our experiments, query expansion is performed with relationship between scientific terminologies with user's intention.

KEY WORDS: Word Embedding, Text Mining, Web Technology, Big Data, Information Retrieval

1 INTRODUCTION

WITH the development of the web, the amount of information is rapidly increasing. Then many people have been acquired unnecessary information against their purposes (Beyer and Lancy, 2012). In addition, the increasing the volume of information is increasing more and more because people actively participate in production and distribution as well as information consumption with changes in the web paradigm (Kim and Rho, 2015). Big data refers to large, complex data that cannot be processed in the conventional way (Kwon et al., 2014). However, the meaning of Big Data is not explicitly defined, but it is known to have three characteristics (The Three V's). For example, Gartner, Inc. described Big Data as:

"Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

On the other hand, big data is defined as structured or unstructured data that cannot be collected, stored, retrieved, analyzed, or visualized by existing methods or tools (Xu et al., 2014). It is difficult to handle with conventional database systems because of the huge amount of processing. In addition, a new paradigm for big data processing should be promised because of the need for real-time data processing and unstructured data processing. Generally, the types of information retrieved from the web are divided according to purpose, but are not provided in a form suitable for each individual. Therefore, it is very important to provide information that is relevant to the user's needs from irrelevant issues in amount of information. Providing personalized information in this way can provide an advertising-based revenue model in addition to providing information itself (Nguyen et al., 2015).

On the other hand, it is not easy to analyze the meaning of contents due to the complexity of language structure in Korean. When a word has ambiguity with multiple meanings, analytic processing is very difficult. Therefore, many studies of text mining try to understand the meaning by the use of words through word sense disambiguation (Han et al., 2018)(Agirre, E. and P. Edmonds,, 2007). Keyword can represent the meaning of a document among the many words that make up a documents. Terminology is a word that is closely related to the content or subject of document. Understanding terms refers to summarization, information retrieval, classification, clustering, it can be applied to various application services [3]. Furthermore, various methods can be used for terminology extraction and statistical analysis with linguistic methods, and machine learning based methods [4]. The statistical approach does not require data for learning and can extract terms based on simple statistical information of the words in the document. Typical methods are n-gram, TF-IDF and co-occurrence based approach [5]. The linguistic

CONTACT Mucheol Kim mucheol.kim@gmail.com

method utilizes the linguistic features that words use in sentences or documents, such as parts of speech, lexical analysis, and parsing. The machine learning method learns how to extract terms from learning data and extracts terms. Recently, the LDA method for finding the subject of a document based on the distribution of words has been utilized, and Word2Vec methods for making and comparing vector information about words through learning of documents have been studied.

In this paper, we propose a method to analyze scientific terminologies with word embedding approach. In particular, we collect scientific data in the BT(Bio Technology) field and analyze the relationship between the scientific terminologies for enhancing knowledge representation. In this process, Word Embedding method is applied to query expansion. As a result, it was confirmed that the query expansion can be performed in accordance with the user's intention.

In this paper, Section 2 presents related works, and section 3 describes the proposed word embedding with science data. In Section 4, experimental analysis and concluding remarks in Section 5.

2 RELATED WORK

THERE are many researches which are recommendation system and text mining with social and science data.

Amount of researches are focused on social data (i.e. social news, SNS) analytics with typical algorithms(Kaur & Gupta, 2010)(Zhang, 2008). They could recommend news items to users according to the content of the news items, collaborative filtering and so on. Therefore, the main difference between content-based recommenders and collaborative filtering recommenders lies in the focus on relationship between user-content similarities.

Some researchers performed typical researches for Korean word sense disambiguation. (Lee et al., 2000) constructed Korean WordNet from pre-existing lexical resources. (Kang et al., 2017) proposed a word sense disambiguation method using word embedding. (Han et al., 2018) also suggested Korean word sense disambiguation with unsupervised learning for a Korean lexical semantic network.

(Kim, 2016) proposed the scientific issue tracking with R&D data such as project reports, research papers, and patents. He analyzed the semantic relationship between terminologies with similarity. (Xu et al, 2016) suggested the matchmaking algorithm for recommending candidates of the research projects.

On the other hand, there are amount of topic analysis in order to apply the application related to the information retrieval and recommender system. LDA (Latent Dirichlet Allocation) (Blei, 2012) which is the issue extracting algorithm based on the probability model is used for extracting topics from social data and news data (Jung et al., 2013)(Kim et al., 2018).

3 WORD EMBEDDING BASED KNOWLEDGE REPRESENTATION APPROACH

IN this chapter, we proposed word embedding based knowledge representation approach with extracting relationship between scientific terminologies. It is separated to extracting similarity between terms and inferring the relationship with word embedding. Then our research could suggest a methodology that can provide users with the results of query expansion among R&D processes which are research paper, patent, project.

3.1 TF-IDF weighting for topic analysis

The TF-IDF (Term Frequency-Inverse Document Frequency) is a typical analysis method for analyzing document characteristics in a vector space model. The TF-IDF function weights each vector component of each document based on the following criteria. (Soucy et al., 2005).

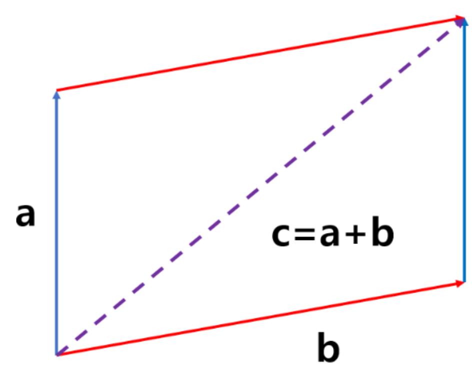


Figure 1. The example of addition of both vectors

First, we identify the characteristics of the document using the frequency of the word. It means that the frequency of the document indicates the importance of the word. On the other hand, IDF means the frequency of occurrence of words between documents, and as the occurrence frequency of words between documents decreases, the uniqueness of words increases. Conversely, as the number of documents containing a word increases, it is judged to be a general-purpose word, and its importance decreases.

In this paper we assumed that each scientific documents are consisted of typical terms which represents the identity for each document (Equation 1).

$$a_i = \{t_1, t_2, t_3, \dots, t_n\} \quad (1)$$

TF means the frequency in each documents, then we calculated the generalized term frequency with maximum number of frequency in each documents.

$$TF(t_n, a_i) = \frac{f_{t_n,a_i}}{\max(f_{t_n,d}: t_n \in a_i)} \quad (2)$$

We also generated the IDF with traditional methodology which is counting the frequency of documents for each terms.

$$IDF(t_n, D) = \log \frac{N}{|\{a_i \in D: t_n \in a_i\}|} \quad (3)$$

As a result, we could deduct the TFIDF value with the harmony of TF and IDF for evaluating the term weight.

$$TFIDF(t_n, a_i, D) = TF(t_n, a_i) \times IDF(t_n, D) \quad (4)$$

3.2 Extracting Relationship with Word Embedding

Word2Vec is a tool that can efficiently estimate the meaning of words in vector space[Milokolov et al., 2013]. In order to estimate words in vector space, multi-layer neural network learning is performed using CBOW (Continuous Bag Of Word) and Skip-gram method, it is possible to compare the relations between words existing in the learned documents in cosine similarity, and to relate words having high relevance such as homonyms.

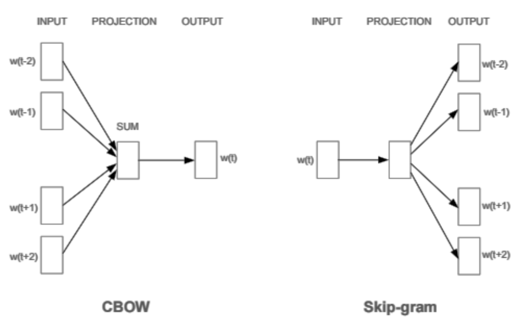


Figure 2. Word2Vec Model for analysing Word Embedding

The frequency of coincidence is traditionally used to see the relevance of words. The relationship between words can be confirmed based on the statistical information of the words simultaneously appearing in the same document. In other words, many words in the same document have higher relevance than those that are not. This method has the advantage of being able to deduce the relationship between simple and intuitive words.

However, Word2Vec can infer the relation between scientific data by using neural network learning based on CBOW and Skip n-gram. Based on their surrounding words, the neural network generates a unique vector, and the higher the similarity of vectors, the higher the correlation. Since all the words are projected in the vector space, they have the advantage that they can deduce the relation of all technical terms. In this study, Word2Vec was used to implement the Word Embedding based relational model.

The bag-of-words model is one of the most representative expression methods for classifying objects. It is applied to the text mining method to identify features through histograms of images. BoW uses the histogram of words to understand the meaning of information, uses the group relation of words appearing together, and ignores the order (Zhang, et al., 2010). Word embedding is one of the most popular representation of document vocabulary. It could predict context of a word in a document. Word embedding is vector representations of a particular word for generating their contexts.

CBOW is to treat context and from these words, and it could predict the center word with surrounding context (Liu, 2018). The order of the words in context does not affect the prediction. Then two vectors would be produced by calculating probabilities. The error vector for each output layer is produced in the manner as discussed above. However, the error vectors from all output layers are summed up to adjust the weights via backpropagation. This ensures that weight matrix for each output layer remains identical all through training.

$$\mathcal{L}_{CBOW}(D) = \frac{1}{M}\sum_{i=1}^{M} \log p(w_i|w_{cxt}) \quad (5)$$

Skip-gram model focused on the use of target and context words. In this case, the target word is fed at the input, the hidden layer remains the same, and the output layer of the neural network is replicated multiple times to accommodate the chosen number of context words. That is, the model predicts the window of surrounding words using the distance weights between words around the current word.

$$\mathcal{L}_{Skip-gram}(D) = \frac{1}{M}\sum_{i=1}^{M} \sum_{-k\le c\le k, c\ne 0} \log p(w_{i+c}|w_i) \quad (6)$$

The relationship between scientific terminologies is constructed from the dynamically generated science data using the results of the scientific term vector calculated by probability distribution. In this research, we utilized Word2Vec tool to express a scientific data set in a vector space of 200 dimensions. As a result, word embedding is performed by analyzing the relationship between documents and terms. The results are applied to query expansion of user queries to provide relevant science terms.

In this paper, we analyzed user's single and complex query, then recommended top-k results with high relevance. They are including the solution to distinguish between homonyms and synonyms. Homonym means that there are the same term, but they have different meanings. Synonyms have different form, however, same or similar meaning. For example, an apple could be analyzed to a fruit, a company, and adult contents.

Associative queries in the query 'apple' can coexist with 'banana' or 'i-phone'. This can lead to confusion if the user does not have enough background knowledge. Thus, a clustered association search can provide results consistent with the user's intentions.

The proposed clustering approach was performed based on similarity between scientific terminologies. The cluster head could be selected with the highest similarity value in a cluster. Then the size of the cluster changes according to the threshold value. After that, the similarity distance between the words in the cluster is evaluated, and the most central word is changed to the head.

Finally, and the similarity-based clustering of the scientific data is repeated until there is no further change. As a result, high-relevance words are suggested as candidate candidates for query expansion, and semantic information provision as well as scientific terminology can be utilized for users.

4 EXPERIMENTS AND ANALYSIS

4.1 Experimental Setup

IN the proposed approach, we constructed data set from Open API provided by NDSL. We collected data on research papers, patents, and research reports which are based on essential keywords in the BT field (Table 1).

Table 1. Query for Crawling science data in BT

	Query for Crawling Science Data in BT
Query Sets	Senior(노인), Rural(농촌), Nursing(간호직), Public(공공), Mental(정신), Environment(환경), Oral(구강), Safety(안전), Medical Law(의료법), Home(재가), School(학교), Medical reform(의료개혁), Medical Technology(의료기술), Clinic(진료소), Doctor(의사), Industrial safety(산업안전), Nurse(간호사), Medical Information(의료정보), Medical Policy(의료정책), Welfare(복지)

In this study, the scientific terminology is limited to the nouns in the scientific document, and the pre-processing is performed using the NLP library (Jeon, 2012). In addition, the TF-IDF score was calculated for the collected scientific data, and an idiomatic dictionary was constructed by manual operation. In addition, morpheme analysis errors and nouns that are not related to scientific terminology are removed. The current status of the data constructed in this study is shown in Table 2.

Table 2. The number of collected Scientific Data

	Paper	Patent	Report
Documents	32,173	34,196	51.388
Terminology	22,853	33,532	75,746
Stop Word	3,961	2,234	1,744

4.2 Experimental Results

(Table 3) is the clustering result of the related terms for the user's query '의사' which is homonym terminology. The first group finds that the terms are related to the hospital doctor. On the other hand, the second and third groups can confirm that words related to the decision-making doctrine mean by intention or communication. A user may refer to a term '의사' as a doctor or as an intention. Therefore, the proposed approach supported query expansion based on clustering terms. As a result, each cluster represents the interrelated words of terms having different meanings, and suggests various methodologies associated with domains.

Table 3. Analytic Results with query '의사'

Cluster 1 with Query 'Doctor(의사)'

Keyword	Score
Patient (환자)	0.4899
Medical Team (의료진)	0.4407
Treatment (진료)	0.4277
Name of a disease (병명)	0.3978
Nurse (간호사)	0.3838
Checkup (검사명)	0.3503
Guardian (보호자)	0.3480
Prescription (처방)	0.3366

Cluster 2 with Query 'Intention(의사)'

Keyword	Score
Multicomponent (다요소)	0.4836
Determinant (결정자)	0.4656
Arbiter (결정권자)	0.4471
Accuracy (정확)	0.4201
Multi criteria (다기준)	0.4050
Manager (경영자)	0.3925
Dialogist (대화자)	0.3886
Communication (의사소통)	0.3878
Analyst (분석가)	0.3676
Rationality (합리)	0.3636

Cluster 3 with Query 'Expression(의사)'

Keyword	Score
Confidence (자신)	0.4831
Decision Unit (결정부)	0.4409
Choice (선택)	0.4187
Input (입력)	0.4011
Request (의뢰서)	0.3903
Later (나중)	0.3893
Confusion (당황)	0.3853
Explanation (설명)	0.3804
adversary (상대방)	0.3727
Feasible (여부)	0.3716

(Table 4) shows the result of compound query of 'doctor' and 'hospital', which is one of the key terms among the clusters. As the preceding single query('의사') has two meanings, it can be confirmed that it is clustered into 'decision making' and 'doctor of hospital'. However, if you enter the word hospital together, the meaning of the doctor becomes clear to the hospital doctor, and you can see that the related terms related to doctors and hospitals are retrieved. In this case, it can be seen that the term 'doctor' is used in a detailed form divided into roles and occupations. In the first cluster, a list of terms related to the treatment was generated. In the second cluster, domain-related terms in the hospital, such as nurses and medical care, are presented.

Table 4. **Analytic Results with compound query**

Cluster 1

Keyword	Similarity
Treatment (진료)	0.6690
Medical Team (의료진)	0.5895
Medical (의료)	0.5282
Rounds (회진)	0.4457
Prescription (처방)	0.4438
Examination (진찰)	0.4175
Second-visit (재진)	0.3981

Cluster 2

Keyword	Similarity
Nurse (간호사)	0.5678
Hospital Administration (원무)	0.5292
Specialist (전문의)	0.5248
Pharmacy (약국)	0.4861
Emergency (응급실)	0.4753
Internal Medicine (내과)	0.4726
Pediatric (소아과)	0.4717
Request (의뢰서)	0.4672
Family Doctor (주치의)	0.4654
Opening Doctor (개원의)	0.4594

One of the results of the proposed approach are summarized as follows. (Table 5). The results show that these results distinguish the features of the query 'Senior(노인)' in detail. Some scientific terminologies indicates the characteristics of senior, as well as facility information for senior, and the term of diseases.

Table 5. **Query expansion with 'Senior'(노인) in BT Patents**

Relationship between terms with Senior('노인')	Similarity
Senior-Weakness('허약')	0.574
Senior - Senior Citizen center('경로당')	0.569
Senior – Life('삶')	0.515
Senior – Treatment('요양')	0.513
Senior – remarriage('재가')	0.500
Senior – Subject('대상자')	0.479
Senior – Depression('우울')	0.472
Senior – Body('신체')	0.470
Senior – Residence('거주')	0.452
Senior – Women('여성')	0.449
Senior - Elderly ('고령')	0.444
Senior – Mild('경증')	0.440
Senior - Welfare Center('복지관')	0.438
Senior – melancholy('우울감')	0.436
Senior - Old age ('노년기')	0.430

The results were used to identify the characteristics of compound queries and to derive similarity evaluation results. The first example shows a Senior-Senior Citizen Center(노인-경로당). The results showed that scientific terminologies related to facilities for seniors were highly prioritized, and dance and yoga related to hobbies also appeared. On the other hand, in the case of Senior-Depressed(노인-우울), it was found that the causes of depression such as Laugh, weakness, melancholy, loneliness, remarriage.

Table 6. **Comparison with two different Compound Query**

Senior-Senior Citizen Center(노인-경로당)

Weakness(허약)	1.135
residence (거주)	0.993
hall (회관)	0.988
Sanatorium(요양원)	0.981
women (여자)	0.953
community relief center (복지관)	0.951
Senior Club(노인정)	0.939
life (삶)	0.938
Dance(춤)	0.931
Yoga(요가)	0.925

Senior-Depressed(노인-우울)

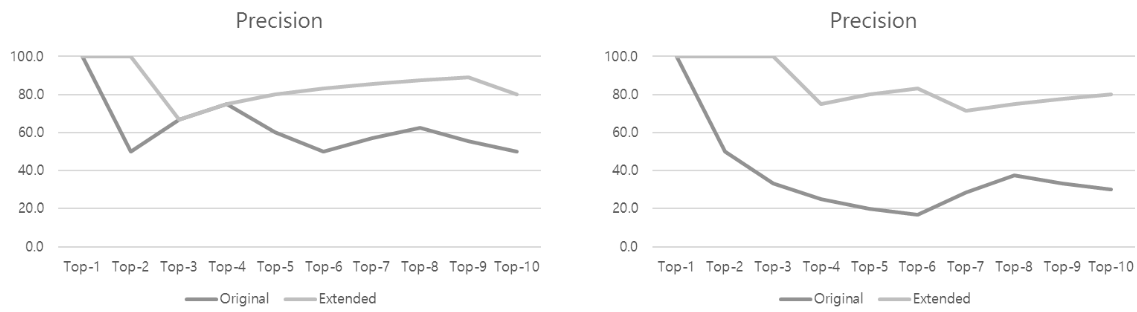
Melacholy(우울감)	1.063
Weakness(허약)	1.052
Senior Citizen Center (경로당)	0.982
Life(삶)	0.981
Loneliness(고독)	0.951
Laugh(웃음)	0.935
Body(신체)	0.934
Respect(존중감)	0.852
Women(여성)	0.840
Remarriage(재가)	0.839

4.3 *Experimental Analysis*

In this section, we analyze the scientific literature by using the proposed research method for deriving the major keywords and extended keywords in the BT field, then evaluate the recommendation results. We evaluated with comparing the top-k recommended results of the query expansion using the proposed word embedding method with the results from existing search single word search. As a result, the accuracy of Top-10 was improved from 0.5 to 0.8 and from 0.3 to 0.8 by 160% and 267%, respectively. (In the experiment of this study, the method of evaluation of recall rate is not suitable, however.) Our results can confirm that the result of the query expansion affects the user's intention by analyzing the semantic relation and it is found that it helps to maintain the recommendation quality even though the number of results to be provided increases . Furthermore, we could suggest the personalized recommendation with users' preferences for extracting relationship between keywords.

5 CONCLUSIONS

RECENTLY, as the amount of information increases dramatically, many people want to acquire the information they want. Furthermore they might getting the information from a reliable source in a timely manner. However, due to the characteristics of big data such as their quantitative volumes and diversity, unnecessary information is increasingly encountered. In this paper, we propose an approach to analyze the scientific terminologies based on word embedding approach. In our experiments, we



(a) Recommdation with query 'Doctor'

(b) Recommdation with query 'Senior'

Figure 3. Precision Results with the proposed approach

proposed a methodology of query expansion by collecting data on papers, patents, and research reports in the field of Bio Technology. As a result, it was confirmed that the query expansion can be performed in accordance with the user's intention. Future research should analyze the flow of scientific data that changes with time, and study how to provide information to cope with social issues. Future works will focus on providing personalized recommendation through text mining based big data analysis and digital curation.

6 ACKNOWLEDGEMENT

THIS research was supported by the Chung-Ang University Research Scholarship Grants in 2019.

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP; Ministry of Science, ICT & Future Planning) (No. 2017R1C1B5018048).

7 REFERENCES

E. Agirre. and P. Edmonds, Word Sense Disambiguation: Algorithms and Applications, Springer, 2007.

M. Beyer., D. Laney. The importance of 'big data': A definition, Gartner. Retrieved from http://www.gartner.com/resId=2057415, 2012.

D. M. Blei (2012), Probabilistic topic models, Communications of the ACM, Vol. 55, No.4, pp. 77-84.

K. Han, S. Nam, J. Kim, Y. Hahm, K. S. Choi (2018). Unsupervised Korean Word Sense Disambiguation using CoreNet. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

H. Jeon (2012) "KoNLP: Korean NLP package." R package version 0.76 8.

D. Jung, J. Kim, K. Kim, J. Hur, B. Ohn, M. Kang (2013), A Proposal of a Keyword Extraction System for Detecting Social Issues, Journal of intelligence and information systems 19(3), pp. 1-23.

M. Y. Kang, B. Kim, J. S. Lee (2017). Word Sense Disambiguation Using Embedded Word Space. Journal of Computing Science and Engineering, 11(1), 32-38.

J. Kaur, V. Gupta (2010) Effective approaches for extraction of keywords. Journal of Computer Science, 7.6: 144-148.

M. Kim, B. B. Gupta, S. Rho (2018). Crowdsourcing based scientific issue tracking with topic analysis. Applied Soft Computing, 66, 506-511.

M. Kim, and S. Rho (2015), Dynamic knowledge management from multiple sources in crowdsourcing environments, New Review of Hypermedia and Multimedia 21, no. 3-4: pp.199-211.

O. Kwon, N. Lee, B. Shin. (2014), Data quality management, data usage experience and acquisition intention of big data analytics International Journal of Information Management, 34 (3), pp. 387-394

B. Liu (2018). Text sentiment analysis based on CBOW model and deep learning in big data environment. Journal of Ambient Intelligence and Humanized Computing, 1-8.

C. Lee, G. Lee, S. J. Yun (2000). Automatic WordNet mapping using word sense disambiguation. In Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13 (pp. 142-147). Association for Computational Linguistics.

T. Mikolov, K. Chen, G. Corrado, J. Dean (2013), Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

NDSL, http://www.ndsl.kr

P. Nguyen, P. Tomeo, T. Di Noia, E. Di Sciascio (2015). An evaluation of SimRank and Personalized PageRank to build a recommender system for the Web of Data. In Proceedings of the

24th International Conference on World Wide Web (pp. 1477-1482). ACM.

I. Park, S. Kim (2015), A Case Study on the National R&D Projects Applying Project Management Methodology, Management education review, 30(3), pp.455-486.

P. Soucy and G. W. Mineau (2005), "Beyond TFIDF weighting for text categorization in the vector space model." In IJCAI, vol. 5, pp. 1130-1135.

X. Wu, X. Zhu, G. Q. Wu, W. Ding (2014), Data mining with big data. Knowledge and Data Engineering, IEEE Transactions on, 26(1), pp. 97-107.

Y. Zhang, R. Jin, Z. H. Zhou (2010). Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics, 1(1-4), 43-52.

C. ZHANG (2008), Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems, 4.3: 1169-1180.

8 NOTES ON CONTRIBUTORS

Mucheol Kim is a faculty in the School of Computer Science and Engineering at Chung-Ang University. He received the BS, MS, Ph.D. degrees from the school of Computer Science and Engineering at Chung-Ang University, Seoul, Korea in 2005, 2007 and 2012, respectively. He was an assistant professor in a department of computer & software engineering at Wonkwang University(2017-2018). In 2014-2016, he was an assistant professor of Department of Media Software at Sungkyul University, Korea. In 2011-2014, he had been working as a Senior Researcher in Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea. His research interests include Information Retrieval, Web Technology, Social Networks and Wireless Sensor Networks.

Junho Kim is a master course student in the department of Computer Science and Engineering at Chung-Ang University. He received the BS from the department of Computer Science and Engineering at Wonkwang University, Iksan, Korea in 2018. His research interests include Big Data, Blockchain, data mining.
kjhcau@dilab.cau.ac.kr

Mincheol Shin is a master course student in the department of Computer Science and Engineering at Chung-Ang University. He received the BS from the department of Computer Science and Engineering at Wonkwang University, Iksan, Korea in 2019. His research interests include Big Data, Blockchain, data mining.
mcsin1648@dilab.cau.ac.kr