Indonesian Journal of Electrical Engineering and Computer Science
Vol. 19, No. 3, September 2020, pp. 1533~1539
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v19.i3.pp1533-1539

Main keyword comparison based on document analysis system

Jongwon Lee, Jaeseung Lee, Hoekyung Jung
Department of Computer Engineering, Paichai University, South Korea

Article Info

Article history:

Received Feb 10, 2020
Revised Mar 1, 2020
Accepted Mar 15, 2020

Keywords:

Deduplication
Document analysis
Keyword
Paragraph extraction
Sequence maintenance

ABSTRACT

Existing document analysis systems list words in the document using a morpheme analyzer. Such a structural feature is difficult to help users to understand the document. To understand a document, you need to analyze the keyword in the document and extract the paragraphs including the keyword. The proposed system retrieves keywords from documents written in XML format, extracts them, and displays them to the user. In addition, it extracts the paragraphs including the keyword entered by the user and maintains paragraph sequence and delete for duplicate paragraphs. Then, the frequency and weight of the keyword are calculated, and the number of paragraphs is reduced by removing the paragraphs including the keyword having a weight less than other keywords weighed. This method may reduce the time and effort required for the user to understand the document as compared to the existing document analysis systems.

Corresponding Author:

Hoekyung Jung,
Department of Computer Engineering,
Paichai University,
155-40 Baejae-ro, Seo-gu, Daejeon, South Korea.
Email: hkjung@pcu.ac.kr

1. INTRODUCTION

Most existing document analysis systems use morphological analysis. Those systems use the functions of the morpheme analyzer. Also, the systems show word list. The word list consists of words in the document [1-3]. Other systems perform the function of searching for documents or paragraphs including keywords that the user inputs [4]. However, existing systems do not understand documents efficiently because they just show keywords or paragraphs [5-7]. In order to efficiently understand a document, if a user inputs a keyword then, the system must search for paragraphs including that keyword and extract them [8], [9]. Also, extracted paragraphs are analyzed to form important paragraphs and displayed to the user [10-12].

In this paper, we propose a system for extracting important paragraphs. The system helps user to efficiently analyze XML document type reports and articles [13]. It also maintains the sequence of the paragraphs and removes duplicate paragraphs. Then, the weight of the keyword is calculated, and the system removes paragraphs including keyword of lowest weight [14-16]. This function increases the compression rate. This suggests that the proposed system can shorten the time required to analyze documents compared to existing systems [17].

2. SYSTEM DESIGN

This section describes the design of the proposed system. The system is designed in three hierarchical structures [18-20]. The system was implemented in Java, which makes it possible to run in various environments without depending on OS. Figure 1 shows the structure of the system to meet the requirements and Figure 2 shows the flow of the system [21].

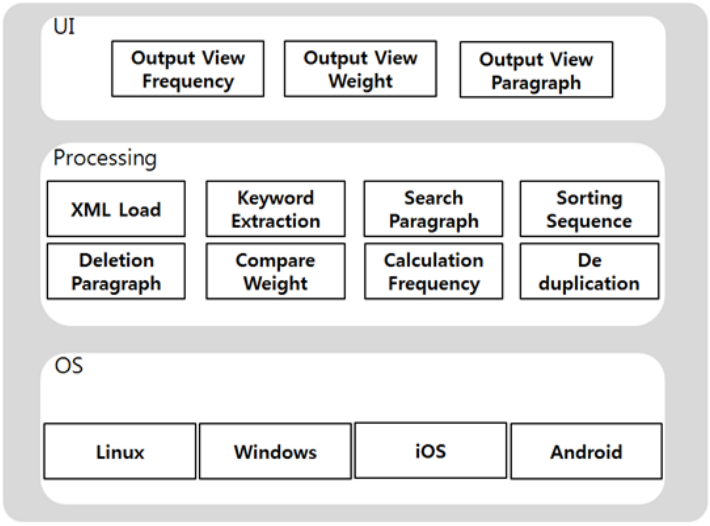

Figure 1. System configuration

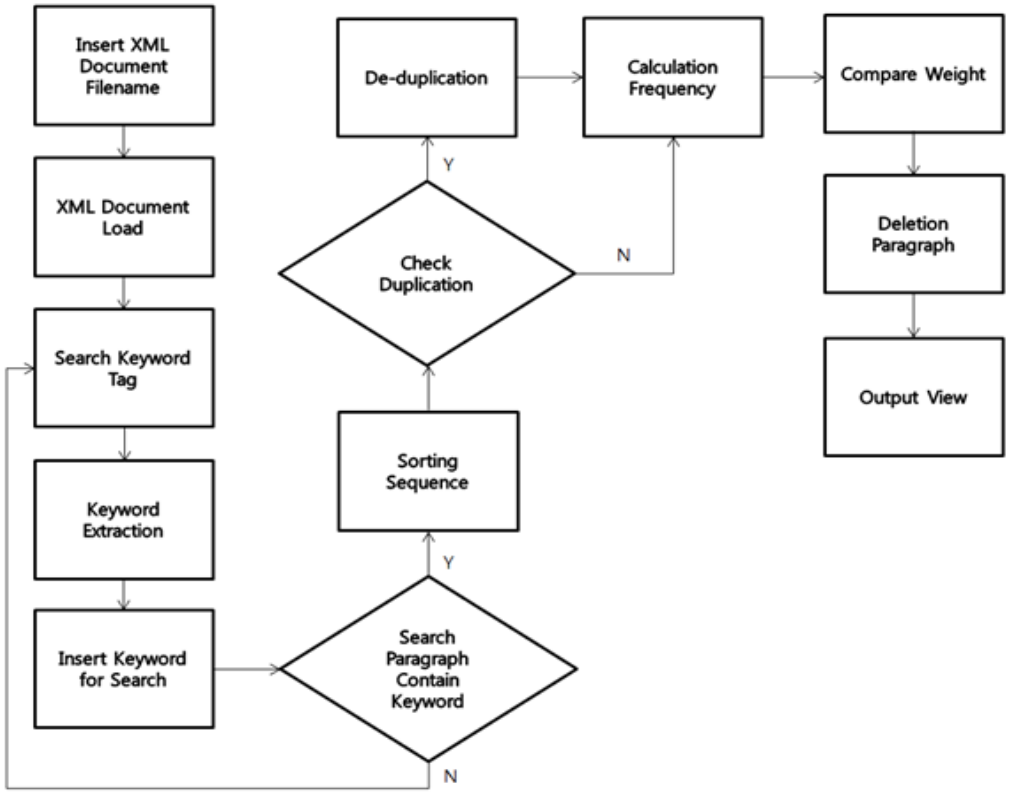

Figure 2. System processing

The functions required for system design are as follows.

a) Function to load user-entered documents.
b) Function to search and extract keyword from a document.
c) Function to search and extract paragraphs including keyword.
d) Function to maintain and sort the sequence of extracted paragraphs.
e) Function to check for duplicates of extracted paragraphs.
f) Function to calculate the frequency of keyword and compare it to weight of other
g) Function to remove paragraphs including keyword of lowest weight.

The System Processor retrieves keyword tags, extracts keyword, and displays the keyword. Subsequently, the user inputs a keyword that they want to search. When the user inputs a keyword, the system searches and extracts the paragraphs including the keyword. Then the system sorts the extracted paragraphs in their original order and check for duplicate paragraphs. If a redundant paragraph is detected, the system will deduplicate repetitive paragraphs. The system then calculates the frequency and weight of the keyword. Also, the system removes paragraphs including keyword of lowest weight [22-25]. The system displays the keyword, weight of keyword, and paragraphs. It can receive refined information more than existing systems based on morpheme analysis. In addition, the system extracting main paragraphs are considered to be able to shorten the time required for document comprehension.

3. SYSTEM IMPLEMENTATION

This chapter describes the implementation and efficiency of the proposed system. Implementation and experiments were using a PC. The PC is OS Windows, CPU - Intel i5-4690, RAM - 8. When the system starts, the user inputs the file name user wants to analyze. Then the file loads using function of the FileInputStream class. And, function of Buffer class is utilized, that reads the contents of the file. Figure 3 shows the flow of the function.

The system searches for the document with the file name entered by the user. Then system searches the keyword tag and extracts the keyword, then system displays it to the user. Figure 4 shows the flow of the function. When the keyword extraction is completed, the user inputs a keyword. Then the system searches for paragraphs that contain the keyword that the user entered and extracts them. Figure 5 shows the function flow. After completing paragraphs search including keywords, the system counts the number of paragraphs. Figure 6 shows the flow of the function to count the number of paragraphs including keyword. The system does sequence maintenance function and deduplication function of the paragraph. Then, the system calculates number of paragraphs and keyword weight. Figure 7 shows the sequence maintenance and deduplication flow.

Figure 3. XML document file open flowchart

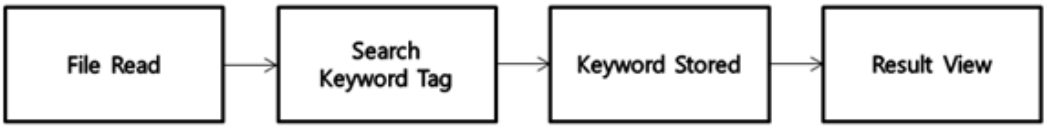

Figure 4. Keyword extraction flowchart

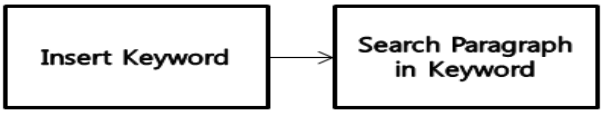

Figure 5. Search paragraph flowchart

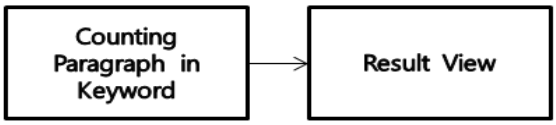

Figure 6. Counting paragraphs flowchart

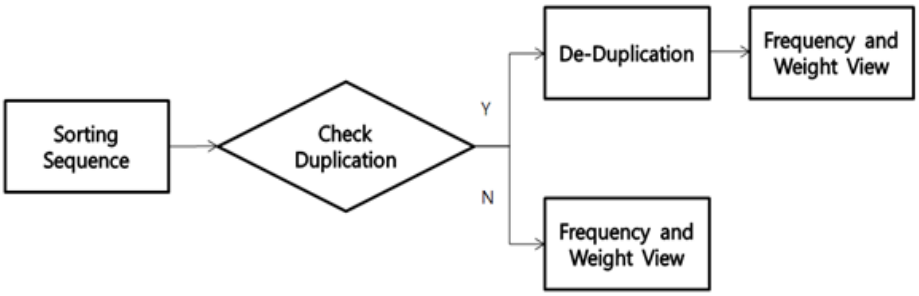

Figure 7. Check duplication flowchart

After sequence maintenance and deduplication, system checks the frequency and weight of keyword. Then, the frequency of the keyword is expressed as a percentage. Next, the system checks for keyword of the lowest frequency, also the system checks for keyword of lowest frequency and the system removes the paragraphs including the keyword of lowest frequency.

If there are paragraphs that contain only a specific keyword but no other keyword, the system removes paragraphs. If there are two or more other keywords which form part of the keyword with lowest weight, then system does not remove such keywords. If a paragraph contains more than one word, it makes it difficult to understand the document. Figure 8 shows the flow of function. This calculates keyword frequency and keyword weight and determines whether to output paragraphs based on keyword frequency.

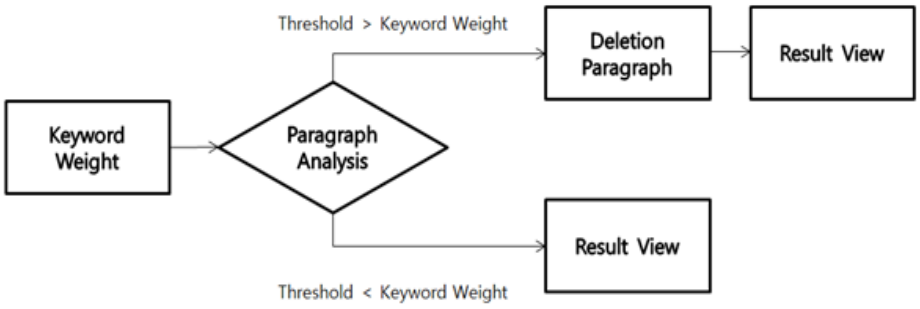

Figure 8. Paragraph analysis flowchart

Most existing document analysis systems were developed based on morpheme analyzer. For this reason, the main purpose of the existing system was to classify the words used in the creation of the document and to check the frequency. In addition, the existing system only checks identity to whether the document includes a keyword. Then user must read all paragraphs in the document. These results a problem of not being able to reduce the time required to understand the document. So, existing systems cannot help users to understand documents because of structural limitations. If the system can extract paragraphs including user inputted keyword then, the system can help to understand the document.

To solve these problems, the proposed system extracts the paragraphs including the keyword entered by the user. Also, the system does sequence maintenance of extracted paragraphs and remove duplicate paragraphs. Then, the frequency of the keyword is checked, and the weight is calculated and displayed to the user. Experiments were conducted with six normalized XML documents. We compare the existing system and proposed system.

Figure 9 shows experiment Test 1. In the first experiment ‘Test 1’, the existing system extracted 108 paragraphs. And proposed system extracted 102 paragraphs and system deletes 6 paragraphs. Also, system compare each section centrality then, the system selects a main section including 10 main paragraphs. Figure 10 shows experiment Test2. In the second experiment ‘Test 2’, the existing system extracted 160 paragraphs. And proposed system extracted 72 paragraphs and system deletes 88 paragraphs. Also, system compares each section centrality then, the system selects a main section including 7 main paragraphs. Figure 11 shows experiment Test 3. In the third experiment ‘Test 3’, the existing system extracted 62

paragraphs. And proposed system extracted 40 paragraphs and system deletes 22 paragraphs. Also, system compares each section centrality then, the system selects a main section including 4 main paragraphs. Figure 12 shows experiment Test 4. In the fourth experiment 'Test 4'. The existing system extracted 29 paragraphs. And proposed system extracted 20 paragraphs and system deletes 9 paragraphs. Also, system compares each section centrality then, the system selects a main section including 2 main paragraphs.

Figure 13 shows experiment Test 5. In the fifth experiment 'Test 5', the existing system extracted 29 paragraphs. And proposed system extracted 26 paragraphs and system deletes 3 paragraphs. Also, system compares each section centrality then, the system selects a main section including 3 main paragraphs. Figure 14 shows experiment Test 6. In the sixth experiment 'Test 6', the existing system extracted 49 paragraphs. And proposed system extracted 42 paragraphs and system deletes 7 paragraphs. Also, system compares each section centrality then, the system selects a main section including 4 main paragraphs.

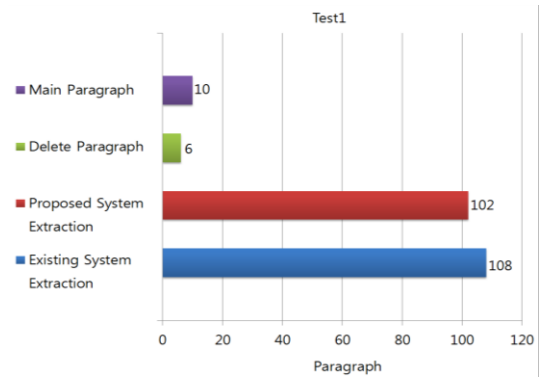

Figure 9. Experiment test 1

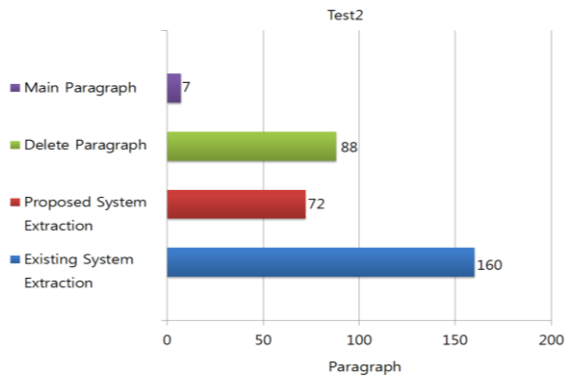

Figure 10. Experiment test 2

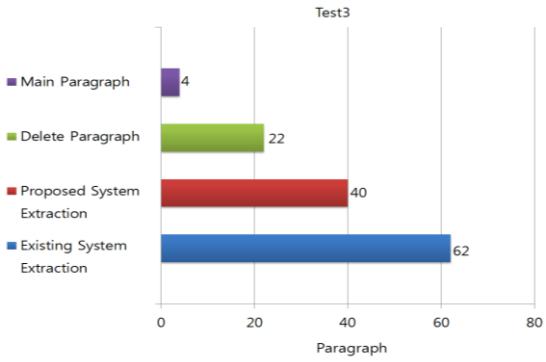

Figure 11. Experiment test 3

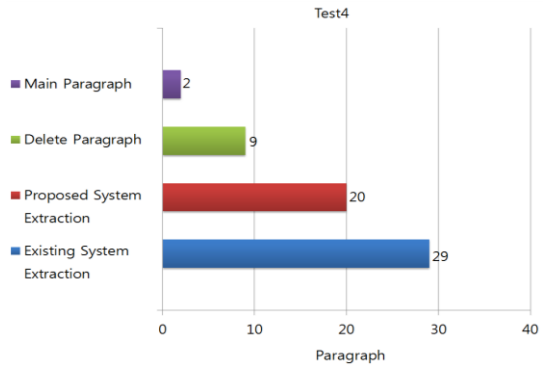

Figure 12. Experiment test 4

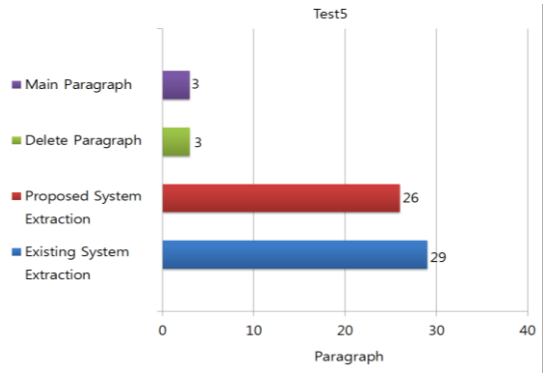

Figure 13. Experiment test 5

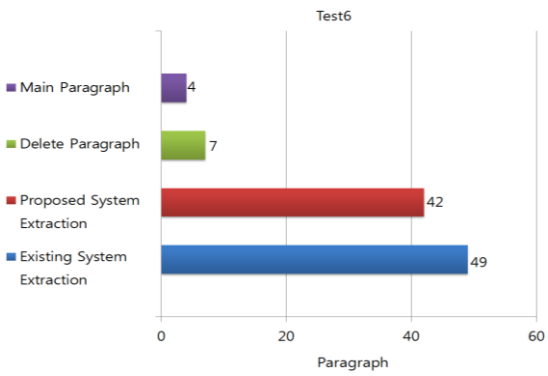

Figure 14. Experiment test 6

Figure 15 shows deleted paragraph for each experiment and Figure 16 shows number of extracted main paragraphs. The Experimental Result proposed that deleted paragraph more than existing system. Also, proposed system compression rate increases for proposed system. Because, the existing system extracted all the paragraphs including the keyword that the user inputs. The existing system extracted all the paragraphs including the keyword that the user inputs. The proposed system is able to sequentially maintain function, deduplication function, and remove paragraphs including keyword of lowest weight. Based on this, the user can see important paragraphs and it is confirmed that the proposed system helps the user to understand the document more efficiently than the traditional system.

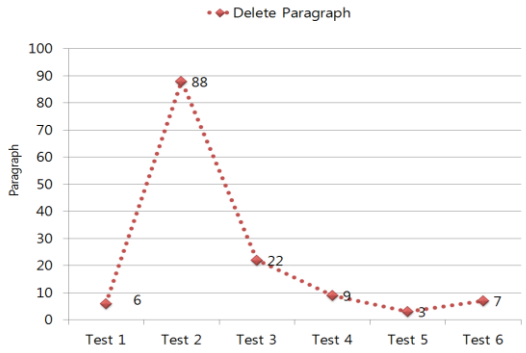

Figure 15. Delete paragraphs

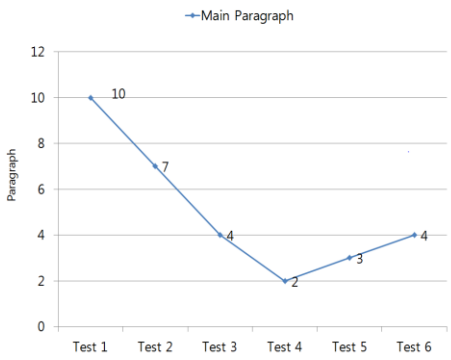

Figure 16. Extracted main paragraphs

4. CONCLUSIONS

The proposed system retrieves the XML document type inputted by the user and then displays the keyword of the document. And, when user inputs the keyword, the system searches and extracts the paragraphs including the keyword. Then, the system does sequence maintenance of paragraphs, and delete duplicate paragraphs if they exist. Then, the system calculates frequency of the keyword and the keyword weight. Subsequently, the system removes the paragraphs including the keyword of lowest weight. If a paragraph includes two or more keywords, the system does not remove the paragraph. Because, doing this breaks the original document context. Also, this situation is same when the system searches only one keyword. Next, the system sorts paragraphs then, displays the paragraphs to the user. So, the system can help users better understand the document. Therefore, the proposed system is more efficient in analyzing documents better than existing system.

ACKNOWLEDGEMENTS

This work was supported by the research grant of Paichai University in 2020.

REFERENCES

[1] J. R. Li, E. H. Lee, and J. H. Lee, "Sequence-to-sequence based Morphological Analysis and Part-Of-Speech Tagging for Korean Language with Convolutional Features," *Journal of Korean Institute of Information Scientists and Engineering*, vol. 44, no. 1, pp. 57-62, Jan 2017.

[2] K. S. Shim, "Cloning of Korean Morphological Analyzers using Pre-analyzed Eojeol Dictionary and Syllable-based Probabilistic Model," *Journal of Korean Institute of Information Scientists and Engineering*, vol. 22, no. 3, pp. 119-126, Mar 2016.

[3] K. S. Shim, "Automatic Word Spacing Using Raw Corpus and a Morphological Analyzer," *Journal of Korean Institute of Information Scientists and Engineering*, vol. 42, no. 1, pp. 68-75, Jan 2015.

[4] J. H. Lee, K. S. Song, J. A. Kang, and J. R. Hwang, "A study on the efficient extraction method of SNS data related to crime risk factor," *Journal of The Korea Society of Computer and Information*, vol. 20, no. 1, pp. 255-263, Jan 2015.

[5] H. Y. Lee, J. S. Lee, B. D. Kang, and S. W. Yang, "Functional Expansion of Morphological Analyzer Based on Longest Phrase Matching For Efficient Korean Parsing," *Journal of Digital Contents Society*, vol. 17, no. 3, pp. 203-210, Jun 2016.

[6] J. Y. Lee, J. H. Lee, and Y. H. Park, "A design and implementation of the management system for number of keyword searching results using Google searching engine," *Journal of The Korea Institute of Information and Communication Engineering*, vol. 20, no. 5, pp. 880-886, May 2016.

[7] J. Y. Lee, J. H. Lee, and Y. H. Park, "Document Classification Model Using Web Documents for Balancing Training Corpus Size per Category," *Journal of Information and Communication Convergence Engineering*, vol. 11, no. 4, pp. 268-273, Dec 2013.

[8] K. Cao, J. W. Lee, and H. K. Jung, "Keyword Analysis Based Document Compression System," *Journal of Information and Communication Convergence Engineering*, vol. 16, no. 1, pp. 48-51, Mar 2018.
[9] H. S. Ha, and B. Y. Hwang, "Keyword Filtering about Disaster and the Method of Detecting Area in Detecting Real-Time Event Using Twitter," *Journal of Korea Information Processing Society*, vol. 5, no. 7, pp. 345-350, Jul 2016.
[10] J. Yim, B. Hwang, "Twitter Based Realtime Event-Location Detector," *Journal of Korea Information Processing Society Transactions on Software and Data Engineering*, vol. 4, no. 8, pp. 301-308, 2015.
[11] X. Zhou, L. Chen, "Event Detection over Twitter Social Media Streams," *The International Journal on Very Large Data Bases*, vol. 23, no. 3, pp. 381-400, Jun 2014.
[12] S. H. Na, J. I. Kim, E. J. Lee, P. K. Kim, "A Study on the Short Text Categorization using SNS Feature Informations," *Journal of Korean Institute of Information Technology*, vol. 14, no. 6, pp. 159-165, Jun 2016.
[13] J. H. Kwon, D. K. Lee, "Social Search Engine using Location based Social Network Service," *Journal of The Korean Institute of Information Technology*, vol. 10, no. 3, pp. 179-187, Mar 2012.
[14] D. W. Kim and M. W. Koo. "Categorization of Korean News Articles Based on Convolutional Neural Network Using Doc2Vec and Word2Vec," *Journal of Korean Institute of Information Scientists and Engineering*, vol. 44. no. 7, pp. 742-747, Jul 2017.
[15] J. M. Kim and J. H. Lee, "Text Document Classification Based on Recurrent Neural Network Using Word2vec," *Journal of Korean Institute of Intelligent Systems*, vol. 2, no. 6, pp. 560-565, Dec. 2017.
[16] R. Li, K. H. Lei, R. Khadiwala, and K. Chang, "TEDAS: a Twitter Based Event Detection and Analysis System," *Proc. of the IEEE 28th International Conference on Data Engineering*, pp. 1273-1276, Apr 2012.
[17] M. Y. Ren and S. J. Kang. "Comparison Between Optimal Features of Korean and Chinese for Text Classification," *Journal of Korean Institute of Intelligent Systems*, vol. 25, no. 4, pp. 386-391, Aug 2015.
[18] J. Shin and C. Ock, "A Stage Transition Model for Korean Part-of-Speech and Homograph Tagging," *Journal of Korean Institute of Information Scientists and Engineers*, vol. 39, no. 11, pp. 889-901, Nov 2012.
[19] I. S. Kang. "A Comparative Study on Using SentiWordNet for English Twitter Sentiment Analysis," *Journal of Korean Institute of Intelligent Systems*, vol. 23, no. 4, pp. 317-324, Aug 2013.
[20] K. R. Kim, D. Y. Lee and H. G. Cho, "Keyword Network Visualization for Text Summarization and Comparative Analysis," *Journal of Korean Institute of Information Scientists and Engineering*, vol. 44, no. 2, pp. 139-147, Feb 2017.
[21] K. B. Lee, J. B. Baik, S. W. Lee, "Estimating a Pleasure-Displeasure Index of Word based on Word Similarity in SNS," *Journal of Korean Institute of Information Scientists and Engineers*, vol. 20, no. 3, pp. 159-164, Mar 2014.
[22] S. J. Choi, J. W. Lee, "A Morphological Analysis Method of Prediction place-Event Performance by Online News Titles," *Journal of Korea Association of Community Welfare Studies*, vol. 21, no. 1, pp. 15-32, Feb 2016.
[23] S. E. Pratama, W. Darmalaksana, D. S. Maylawati, H. Sugilar, T. Mantoro, M. A. Ramdhani, "Weighted inverse document frequency and vector space model for hadith search engine," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 1004-1014, May 2020.
[24] E. Seshatheri, T. Bhuvaneswari, "Effective XQuery keyword using XML query processing," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 450-454, Apr 2019.
[25] N. Kamaruddin, A. W. A. Rahman, R. A. M. Lawi, "Jobseeker-industry matching system using automated keyword selection and visualization approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 3, pp. 1124-1129, Mar 2019.

BIOGRAPHIES OF AUTHORS

Jongwon received the M.S. degree in 2016 and Ph. D. degree in 2019 from the Department of Computer Engineering of Pai Chai University, Korea. His current research interests include multimedia information processing, information retrieval system, and semantic web.

Jaeseung Lee received the M.S. degree in 2019 from the Department of Computer Engineering at Pai Chai University. He is currently a doctoral candidate in the Department of Computer Engineering at Pai Chai University. Since 2016, he has been working as an encryption currency specialist. His current research interests include information processing, IoT, big data, and blockchain.

Hoekyung Jung received the B.S degree in 1987 and Ph. D. degree in 1993 from the Department of Computer Engineering of Kwangwoon University, Korea. From 1994 to 1995, he worked for ETRI as a researcher. Since 1994, he has worked in the department of Computer Engineering at Paichai University, where he now works as a professor. His current research interests include multimedia document architecture modeling, machine learning, IoT, bigdata, and artificial intelligence.