Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 16(1): 48-51, Mar. 2018 Regular paper

Keyword Analysis Based Document Compression System

Kerang Cao[1], **Jongwon Lee**[2], **and Hoekyung Jung**[2*], ***Member*, *KIICE***

[1]Department of Computer Science and Engineering, Shenyang University of Chemical Technology, Shenyang, Liaoning 110041, China
[2]Department of Computer Engineering, Pai Chai University, Daejeon 35345, Korea

Abstract

The traditional documents analysis was centered on words based system was implemented using a morpheme analyzer. These traditional systems can classify used words in the document but, cannot help to user's document understanding or analysis. In this problem solved, System needs extract for most valuable paragraphs what can help to user understanding documents. In this paper, we propose system extracts paragraphs of normalized XML document. User insert to system what filename when wants for analyze XML document. Then, system is search for keyword of the document. And system shows results searched keyword. When user choice and inserts keyword for user wants then, extracting for paragraph including keyword. After extracting paragraph, system operating maintenance paragraph sequence and check duplication. If exist duplication then, system deletes paragraph of duplication. And system informs result to user what counting each keyword frequency and weight to user, sorted paragraphs.

Index Terms: Compression, Document analysis, Keyword, Paragraph extraction

I. INTRODUCTION

As the amount of documents increase so, the importance of techniques for analyzing documents is increasing. Also, the types of documents and the purpose of writing them vary. Therefore, there is a demand for a program for help analyze or understand the document. Most existing document analysis systems are based on morphemes [1–3]. This system shows number of frequency they are used so that the user can know the main words of the document. However, user needed time for understanding document can't decrease.

Other types of systems find the paragraph containing the search term you entered and show it to the user. These types of systems help users understand the document, but they can't reduce the time required to understand the document. In addition, because it extracts a paragraph containing all the keywords entered by the user or extracts all the paragraphs containing each keyword, the accuracy and the compression are low [4–7]. To solve this problem, the system must extract the main paragraphs and display them to the user. The proposed system can reduce the time required for the user to understand the document. The proposed system can reduce the time required for the user understand document. The system searches keyword and displays. Next, the system extracts paragraph containing the keyword. The system possible to delete the low important paragraphs, thereby shortening the time required for the user to understand the document.

II. SYSTEM DESIGN

This section describes the design of proposed system. When user wants to analyze XML document then, the system load the document. Then system is extract keyword of the documents and displays. And user insert keyword then,

Received 04 January 2018, **Revised** 15 March 2018, **Accepted** 19 March 2018
***Corresponding Author** Hoekyung Jung (E-mail: hkjung@pcu.ac.kr, Tel: +82-42-520-5640)
Department of Computer Engineering, Pai Chai University, 155-40, Baejae-ro, Seo-gu, Daejeon 35345, Korea.

Open Access **https://doi.org/10.6109/jicce.2018.16.1.48** print ISSN: 2234-8255 online ISSN: 2234-8883

system extract paragraph containing the keyword. If duplicate paragraph are exist then the system removes duplicate paragraphs. Also maintains the sequence of the paragraphs and displays result to the user. Fig. 1 shows the structure of the system and Fig. 2 shows the flowchart of the system.

Fig. 1 shows the configuration of the system. System retrieves the XML documents and extracts keyword. Then user inserts keyword and the system extract paragraphs including the keyword. And system sorts sequence of paragraphs and check duplication. To do this, the system is designed with three hierarchical structures. The proposed system is implemented in Java. So it can be executed in windows and other develops environments.

Fig. 2 shows the flow of the system. The system starts when the user inserts the file name of XML document. Then, the system searches keyword of document and displays to user. If user inputs the keyword, system has searching and extracts the paragraph including the keyword. When after paragraphs extraction, the system checks duplications paragraph. Next, system calculates frequency of each keyword and the system displays each keyword frequency and percentage, sorted paragraphs.

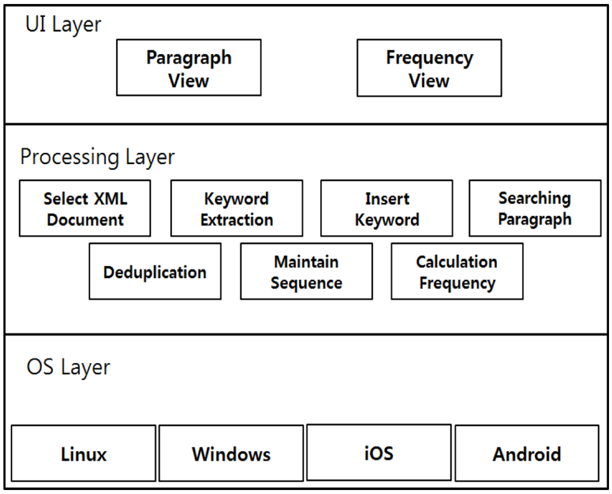


Fig. 1. System architecture.

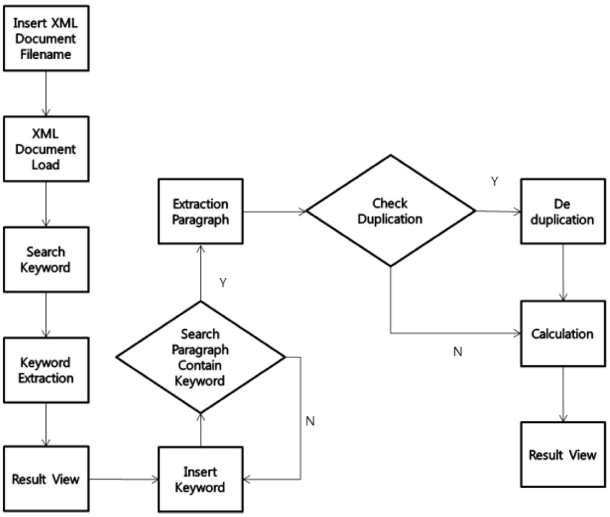


Fig. 2. System flowchart.

III. SYSTEM IMPLEMENTATION

This section describes implementation of proposed system and verify of system efficiency. When the system is started, the user inserts file name of the XML document. Fig. 3 shows the flow of the function.

When the user enters the file name of the XML document, the function of the Java FileInputStream class opens the file and starts to read a line using Java Buffer. System checks next line and if not exist then, finish open file flow. Fig. 4 shows the flow of keyword extraction.

After open the file, system searches keyword of the document. If system finds the keyword then, the system extracts the paragraphs. Also delete stopword in the paragraphs and extracts keyword then finished keyword extraction flow.

And keyword extraction was completed then the extraction

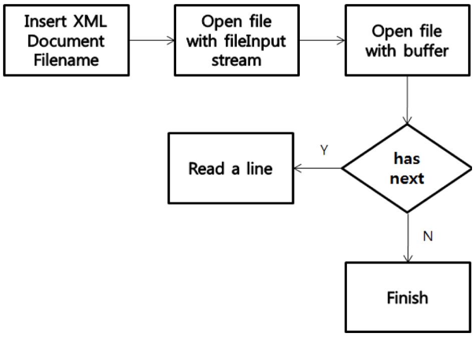


Fig. 3. Open XML document flowchart.

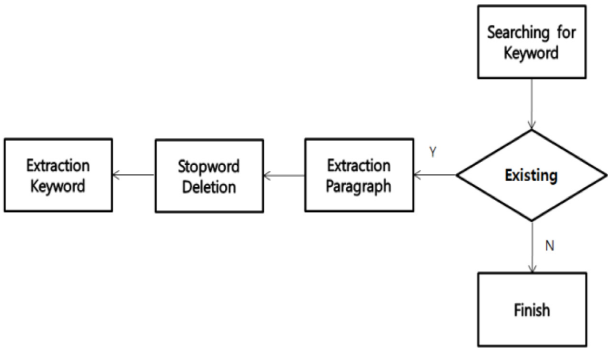


Fig. 4. Extraction keyword flowchart.

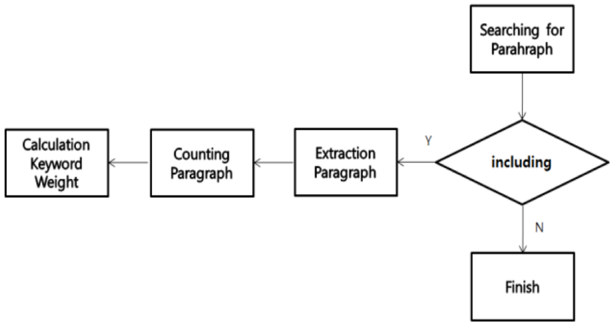


Fig. 5. Extraction paragraph including keyword flowchart.

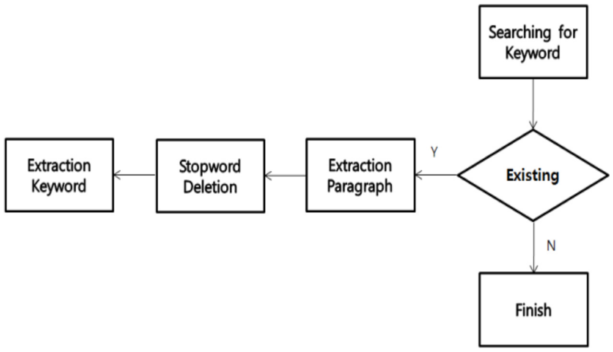


Fig. 6. Check duplication flowchart.

keyword of the XML document is displayed to user. And user inserts keyword of document for user want analysis to system. When the keyword input finish then, the system searches and extracts the paragraphs containing the keyword. Fig. 5 shows the flow of functions that compare with keyword frequency.

The system searches paragraphs including keyword. If such paragraph is including keyword then extracts the paragraph. After system extracts paragraph then, the system shows keyword frequency and weight. Also, the system checks duplication. Fig. 6 shows the flow of de-duplication.

The system checks duplicate and deduplication. And the system informs the number of duplicated paragraphs. In addition, the system results displays of each keyword and each keyword frequency, each keyword weight, percentages. Finally, the system displays of sorted paragraphs.

IV. REVIEW

The traditional extraction paragraph systems are mainly used to classify the words used in document and to confirm the frequency. But, these systems can't reduce the time required to understand of document. To solve this problem, the system extracts the paragraphs including keyword by the user inserts, calculates the frequency and weight of the keyword, displays to user. In addition, duplicate elimination processing is performed on the extracted paragraphs. And the system displays to the user, thereby allowing the user to read the body text including the keyword retrieved by the user. Fig. 7 shows compare the results with the traditional extraction paragraph system in order to verify the efficiency of the proposed system.

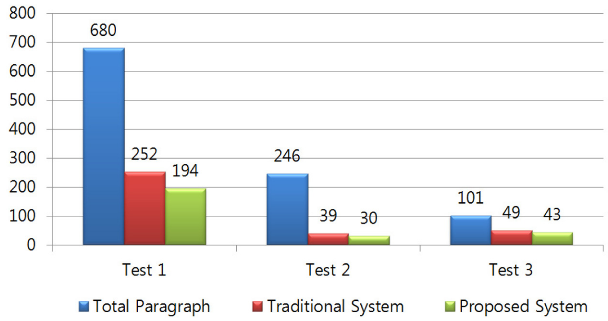


Fig. 7. Test result.

Experiments were conducted three normalized XML documents. We compare to results what traditional extraction paragraph system and proposed system. The y-axis is the number of paragraphs. And in 'Test 1', used No. 1 document, in 'Test 2', used No. 2 document, in 'Test 3', used No. 3 document.

In the first experiment 'Test 1', the number of 680 total paragraphs. Traditional system was extract 252 paragraphs what including keyword. And proposed system was extract 194 paragraphs what including keyword. First experiment results, each system had a compression rate of about 62.95% and 71.48%. So, proposed system is 8.53% higher than the traditional system.

And second experiment 'Test 2', the number of 246 total paragraphs. Traditional system was extract 39 paragraphs what including keyword. And proposed system was extract 30 paragraphs what including keyword. Second experiment results, each system had a compression rate of about 84.15% and 87.81%. So, proposed system is 3.66% higher than the traditional system.

And third experiment 'Test 3', the number of 101 total paragraphs. Traditional system was extract 49 paragraphs what including keyword. And proposed system was extract 43 paragraphs what including keyword. Third experiment results, each system had a compression rate of about 51.49% and 57.43%. So, proposed system is 5.94% higher than the traditional system.

Experiments analysis results, proposed system had an average compression rate 6.04% higher than traditional system. Also, proposed system can compare weight and inform to user what most valuable keyword. So user can analyze document based on main keyword what system shows. Therefore proposed system can help to user who has not all reading paragraphs. First, saved time of analyzing document. Second, user can knew main keyword. We proved the proposed system efficiency better than traditional system.

V. CONCLUSTION

The proposed system is analyzing for normalization XML document. First, user insert XML document file name then system loads the document. And system searches for keyword document. This function finishes then, system displays the keyword to user. Next, when the user inputs the keyword then, the system extracts the paragraphs containing the keyword. After extracting the paragraphs, the system counts each keyword frequency, weight of each keyword are calculated and show. In addition, the system maintenances sequence of paragraphs. So, not damage sequence of paragraphs. And, the system check duplicate. If the system finds duplication paragraph then, the system deletes paragraph.

As a result, the proposed system has higher accuracy and compression rate than traditional systems. It is expected that it will show high efficiency in understanding documents to users in research field analyzing data such as reports or papers in the form of XML documents.

ACKNOWLEDGMENTS

This work was supported by the research grant of Pai Chai University in 2017.

REFERENCES

[1] B. Noh, Z. Xu, J. Lee, D. Park, and Y. Chung, "Keyword network based repercussion effect analysis of foot-and-mouth disease using online news," *Journal of the Korean Institute of Information Technology*, vol. 14, no. 9, pp. 143–152, 2016. DOI: 10.14801/jkiit.2016.14.9.143.

[2] J. Li, E. Lee, and J. H. Lee, "Sequence-to-sequence based morphological analysis and part-of-speech tagging for Korean language with convolutional features," *Journal of the Korean Institute of Information Scientists and Engineering*, vol. 44, no. 1, pp. 57–62, 2017. DOI: 10.5626/JOK.2017.44.1.57.

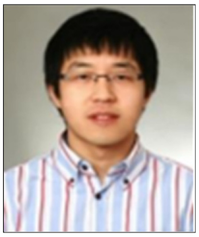
[3] H. Ha and B. Y. Hwang, "Keyword filtering about disaster and the method of detecting area in detecting real-time event using Twitter," *KIPS Transactions on Software and Data Engineering*, vol. 5, no. 7, pp. 345–350, 2016. DOI: 10.3745/KTSDE.2016.5.7.345.

[4] K. S. Shim, "Automatic word spacing using raw corpus and a morphological analyzer," *Journal of the Korean Institute of Information Scientists and Engineering*, vol. 42, no. 1, pp. 68–75, 2015. DOI: 10.5626/JOK.2015.42.1.68.

[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, 2012. DOI: 10.9728/dcs.2016.17.3.203.

[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, 2016. DOI: 10.6109/jkiice.2016.20.5.880.

[7] S. Y. Park, J. Chang, and T. Kihl, "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, 2013. DOI: 10.6109/jicce.2013.11.4.268.



Kerang Cao
received the B.S. degree in 2006 from Computer Science and Application of North Eastern University, China and Ph.D. degree in 2011 from the Department of Computer Engineering of Pai Chai University. From 2016 to the present, he worked for Shenyang University of Chemical Technology as a Lecturer. His current research interests include multimedia document architecture modeling, information processing, IoT, big data, and embedded system.

Jongwon Lee
received the B.S., M.S. degree from the Department of Computer Engineering of Pai Chai University, Korea in 2014 and 2016. He is currently a Doctorate course in Department of Computer Engineering of Pai Chai University. His current research interests include multimedia information processing, information retrieval system, and semantic web.

Hoekyung Jung
received the M.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 Pai Chai University, where he now works as a professor. His current research interests include multimedia document architecture modeling, information processing, information retrieval, and databases.