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

J. Inf. Commun. Converg. Eng. 19(1): 48-53, Mar. 2021 Regular paper

Deep Learning Document Analysis System Based on Keyword Frequency and Section Centrality Analysis

Jongwon Lee[1], Guanchen Wu[2], and Hoekyung Jung[3*], *Member*, *KIICE*

[1]National Science & Technology Information Service, KISTI, Daejeon 34141, Korea
[2]Guizhou Communications Polytechnic, Guiyang 551400, China
[3]Department of Computer Engineering, Pai Chai University, Daejeon 35345 Korea

Abstract

Herein, we propose a document analysis system that analyzes papers or reports transformed into XML(Extensible Markup Language) format. It reads the document specified by the user, extracts keywords from the document, and compares the frequency of keywords to extract the top-three keywords. It maintains the order of the paragraphs containing the keywords and removes duplicated paragraphs. The frequency of the top-three keywords in the extracted paragraphs is re-verified, and the paragraphs are partitioned into 10 sections. Subsequently, the importance of the relevant areas is calculated and compared. By notifying the user of areas with the highest frequency and areas with higher importance than the average frequency, the user can read only the main content without reading all the contents. In addition, the number of paragraphs extracted through the deep learning model and the number of paragraphs in a section of high importance are predicted.

Index Terms: Deep learning, Document analysis, Keyword frequency, Keyword weight, Paragraph extraction

I. INTRODUCTION

Existing document analysis systems have been developed based on morpheme analyzers. A user-specified document is read, whereas words in the document and the frequency of words are sorted. Using morpheme analyzer-based document analysis systems, the user can identify the words used to prepare the analyzed document. However, document analysis systems based on the morpheme analyzer are inefficient because a user can only understand the document when the content of the document has been read. This necessitates the document to be analyzed more comprehensively [1-3].

Herein, we propose a document analysis system that analyzes papers or reports transformed into XML((Extensible Markup Language)) format. The proposed system informs various data to users such that they can understand XML-style papers or reports more efficiently compared with using existing document analysis systems [4-6]. The process of analyzing a document using the system is as follows: First, the system reads the document specified by the user and extracts keywords from the document. Second, the system verifies the frequency of the keywords and uses the top-three keywords. Third, the system extracts the paragraphs containing the top-three keywords and removes duplicated paragraphs. Fourth, the system re-verifies the frequency of the top-three keywords in the extracted paragraphs and partitions the paragraphs into 10 sections. Fifth, the system compares the average frequencies of the top-three keywords in 10 domains and the frequency of a specific domain to calculate the importance of the domain. Sixth, the system notifies the users regarding areas with importance higher than the average importance of the 10 areas. Moreover, if the areas of

Received 22 February 2021, **Revised** 01 March 2021, **Accepted** 02 March 2021
***Corresponding Author** Hoekyung Jung (E-mail: hkjung@pcu.ac.kr, Tel: +82-42-520-5640
Department of Computer Engineering, Pai Chai University, Daejeon 35345, Republic of Korea.

Open Access https://doi.org/10.6109/jicce.2021.19.1.48 print ISSN: 2234-8255 online ISSN: 2234-8883

high importance are adjacent, then the users are notified of the areas [7,8].

By notifying the user of the area with the highest importance, the user can read only the main content without reading all the contents. In addition, using the deep learning model, the number of paragraphs extracted and the number of paragraphs in a section of high importance were investigated. The model analyzes the correlation between the figures and predicts the analysis results of a specific document. Hence, one can analyze the specific document, analyze multiple documents, identify commonalities between the documents, and deliver data that can be clustered to the user.

II. SYSTEM DESIGN

The proposed system analyzes documents in XML format. The reason for specifying the analysis target in XML format is that reports and papers are often stored as XML documents. When the user enters the file name in XML format, the system loads and analyzes the file. The results of the proposed document analysis system are as follows: Users should be informed of the keywords and frequency of the keywords, weights of keywords, paragraphs, importance of domains, and major domains of the document. Therefore, the existing word-based document analysis method and paragraph-based document analysis method are used. Fig. 1 illustrates the structure of the system.

The functions required when designing the proposed system are as follows:

1. Function to load XML format documents entered by users
2. Function to search for keyword tags in a document and extract keywords that are tag values
3. Function to verify keyword frequency and compare it
4. Function to extract paragraphs containing top-three keywords
5. Function to maintain order of extracted paragraphs
6. Function to remove duplicated paragraphs
7. Function to verify frequency of top-three keywords for extracted paragraphs
8. Function to calculate and display weight of top-three keywords
9. Function to divide extracted paragraphs into 10 areas
10. Function to calculate and compare importance of 10 areas
11. Function to inform users of areas of high importance
12. Function to learn number of paragraphs extracted using deep learning model and number of paragraphs in a section of high importance
13. Function to predict number of paragraphs in a section of high importance based on number of paragraphs extracted

To implement the functions, the system was designed in three hierarchies, and a document analysis program was implemented in Java. A deep-learning-based prediction program was implemented in Python. Fig. 2 shows the flow of the system.

When the system begins, the user enters the file name of the XML document to be searched. After reading the XML document of the file name entered by the user, the keyword tag in the document is searched. Subsequently, after loading the keyword, which is the value of the keyword tag, the keyword is shown to the user, and the frequency of the keyword is verified. When the frequency of the keyword is confirmed, the system searches for paragraphs containing the top-three keywords and extracts them. The system maintains the order of the extracted paragraphs and verifies for duplicate paragraphs. If duplicate paragraphs exist, then one paragraph is removed such that only one is printed. In addition, the sys-

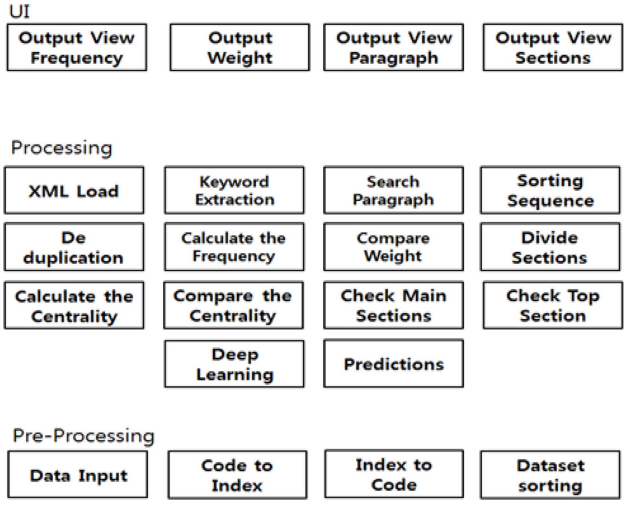


Fig. 1. System architecture.

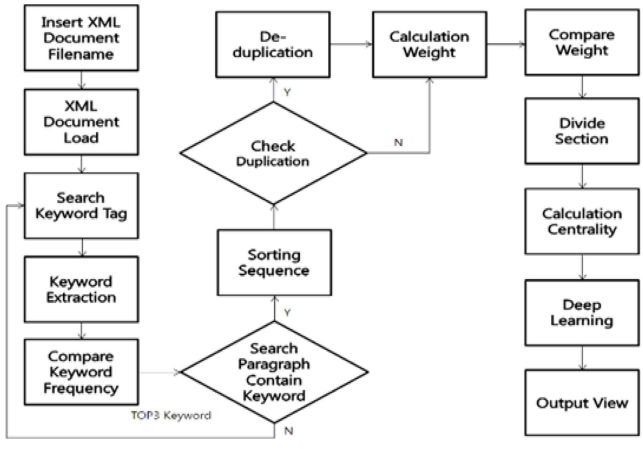


Fig. 2. System flowchart.

tem calculates and compares the weights of the keywords and informs the user of the analysis results. The system partitions the paragraphs into 10 sections and calculates the importance of each section. The importance calculation sums the frequency of keywords and divides them by the number of paragraphs. The importance of 10 areas was summed and divided by 10 to obtain the average value, and the importance and average value of each area were compared. Users are notified of areas of higher importance based on their average values. When the system completes the document analysis, the deep learning model is used to learn the total number of paragraphs extracted by the system and the number of paragraphs in the high-critical area. Subsequently, predictions regarding specific documents are performed. This procedure demonstrates the process by which the number of high-priority paragraphs changes based on the number of paragraphs in a particular document. By notifying the user of these data, the user can organize the relationships between documents or perform clustering.

The proposed system can receive all the functions provided by the existing systems based on the morpheme analyzer. First, the system provides keyword information provided by a word-based analysis system developed based on a morpheme analyzer. Second, the system provides the paragraph information supplied by the paragraph-based analysis system developed based on the morpheme analyzer. Third, the system divides the extracted paragraphs into 10 areas, calculates the importance of each area, and informs the main areas to provide the information necessary for understanding the document. Fourth, the system predicts the analysis result for a specific document based on the results derived by the system using a deep learning model and provides the data to the user. Fifth, the user can organize or cluster the relationships between documents using the information provided by the system.

It is assumed that the proposed system can provide various types of information compared with the existing system as well as extract paragraphs, thereby reducing the time required to understand documents and improve efficiency.

III. SYSTEM IMPLEMENTATION

This section describes the implementation of the proposed system and the verification of its efficiency. The PC used was a Windows operating system CPU(Central Processing Unit)-Intel i5-4690, RAM-8 PC. When the system starts, the user enters a file in XML format for analysis. The system loads the files. In addition, when performing a document import operation, the system identifies the keyword tags, identifies the values of those tags, and extracts them. The extracted keywords are shown to the user, and the frequency is calculated to display the top-three keywords to the user. Subsequently, the document is analyzed based on the top-three keywords. Fig. 3 shows the system output keywords extracted by analyzing a specific document.

Subsequently, the system maintains the order of the paragraphs and removes duplicates. Furthermore, it shows the number of paragraphs that contain the top-three keywords and the weight of the keywords. Fig. 4 shows the screen depicting the results of the system that performs the function.

When the system completes the verification of the frequency and ratio of the top-three keywords, the extracted paragraphs are partitioned into 10 sections. The importance of a domain is to calculate the frequency of the keywords within that domain. In addition, by comparing the importance levels, the area with the highest importance level and areas with an importance level higher than the average of 10 importance levels are notified to the user. In addition, if high-priority areas appear in succession, then the user must be informed to read them. Figs. 5 and 6 show the screens that depict the results after important calculations and comparisons are performed.

Fig. 7 shows the number of average paragraphs from result of 20 experiments. The existing document analysis system and the proposed system extracted 391.48 and 76.8 paragraphs on average, respectively. This indicates a 5:1 ratio, and it was observed that the number of paragraphs suggested by the proposed system to the user was less than that of the existing system. In addition, if the user does not understand the document when reading the paragraphs recommended by the proposed system, then the user should read additional

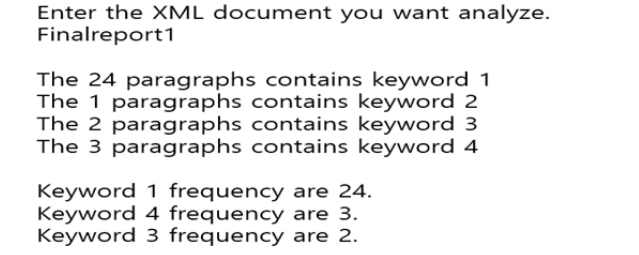


Fig. 3. Screen of keyword extraction.

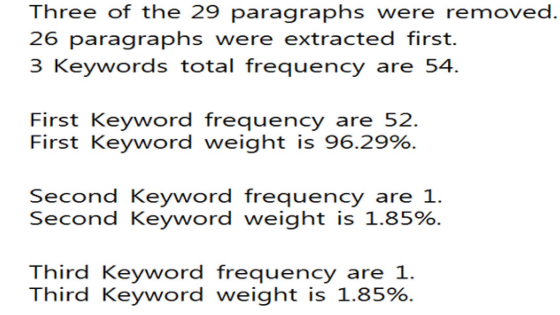


Fig. 4. Screen of comparison results.

```
Section no.1 is divide section.
Section centrality is 2.0
First Keyword frequency are 4
Second Keyword frequency are 0
Third Keyword frequency are 0

Section no.2 is divide section.
Section centrality is 3.0
First Keyword frequency are 5
Second Keyword frequency are 0
Third Keyword frequency are 0

Section no.3 is divide section.
Section centrality is 1.0
First Keyword frequency are 1
Second Keyword frequency are 0
Third Keyword frequency are 0
```

Fig. 5. Screen of centrality output 1.

```
Section no.10 is divide section.
Section centrality is 2.0
First Keyword frequency are 6
Second Keyword frequency are 0
Third Keyword frequency are 0
This section keyword frequency more than average frequency.

Section no.5, no.6, no.7, no.10 are important in this document.
Section no.5, no.6, no.7 are near so, you must read these sections.
Section no.7 centrality is 4.0 then, this section is top section.
```

Fig. 6. Screen of centrality output 2.

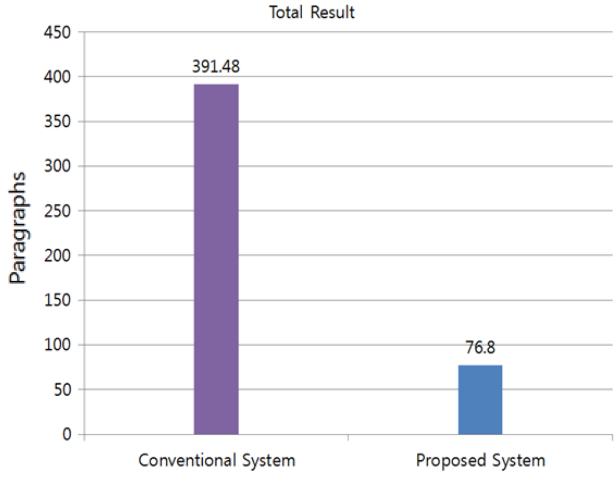
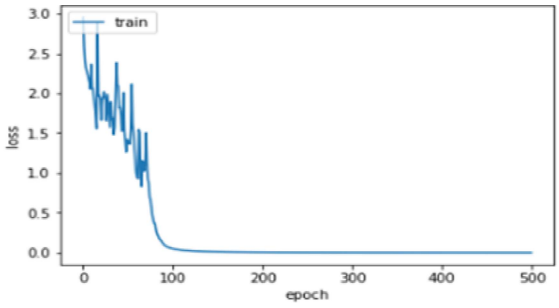


Fig. 7. Graph of total test.

paragraphs from other areas. Nevertheless, the recommended number of paragraphs is less than that of the existing system. However, it can be concluded that the proposed system is highly efficient because the number of paragraphs recommended by the proposed system is significantly less than that by the existing document analysis system.




```
15/15 [==============================] - 0s 21ms/step
acc: 100.00%
```

Fig. 8. Graph of deep learning model accuracy.

Fig. 8 shows the results of applying 20 experiments to a deep learning model.

As a result of using the proposed system, the ratio of all extracted paragraphs to those analyzed as having high importance was 5:1. The deep learning model inserts five data points of the analyzed document before performing predictions and predicts the analysis result for the document to be analyzed later. Beginning from the 76th training, the deep learning model begins to fully match with the predicted value and the actual analysis result. “Loss” on the Y-axis is the loss value, and “epoch” on the X-axis represents the number of repetitions of learning.

IV. ANALYSIS AND DISCUSSIONS

Table 1 lists the results of the sentiment score calculation. This table shows a comparison between the existing and proposed systems.

By analyzing the experimental data 20 times, it was discovered that the compression rate of the proposed system was high in the existing system. Existing document analysis systems developed based on the morpheme analyzer extracted all the paragraphs containing keywords entered by the user. Even if deduplication is performed, a significant amount of content must be read by the user; therefore, it is difficult to reduce the time required for document analysis. Meanwhile, the keywords listed in the document were extracted, the frequency of the keywords was compared, the top-three keywords were identified, the paragraphs containing the keywords were extracted, and then the paragraphs were partitioned into 10 sections. The system is terminated after the areas are calculated and compared and the user notified of areas of high importance.

The proposed system recommends reading fewer paragraphs than the existing document analysis system and notifies users regarding the frequency or weight of the keywords. Hence, the time required for the user to understand the document is less compared with that using the

Table 1. Result of sentiment score calculation

	Conventional System	Proposed System
Automatic Keyword Search	Impossible	Possible
Search Algorithm	Intersection or Union	Union
Deduplication	Certain System Possible	Possible
Maintain Sequence	Certain System Possible	Possible
Keyword Frequency Check	Certain System Progress	Progress
Keyword Weight Calculation	Certain System Progress	Progress
Divide Section of Paragraph	No	Yes
Calculate Centrality of Sections	No	Yes
High Centrality Sections Check and Inform	No	Yes
Deep Learning Based Prediction	No	Yes

existing system. In addition, the proposed system will not alter the content of the document by performing deduplication and the order maintenance of paragraphs. The proposed system provides a higher compression rate and various types of information to the user compared with the existing system and is regarded as useful for document analysis.

V. CONCLUSIONS

The proposed system analyzes documents more efficiently than existing systems by solving the limitations of existing document analysis systems based on the morpheme analyzer and providing the information necessary for users to analyze the documents. Therefore, if it is used for managing and analyzing documents, then the ripple effect will be significant. As a future study, the convenience and efficiency of the system should be verified and UI modification should be conducted.

ACKNOWLEDGEMENTS

This study was supported through a research grant from Pai Chai University in 2021.

REFERENCES

[1] H. S. Lee and J. D. Kim, “A design of similar video recommendation system using extracted words in big data cluster,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 24, no. 2, pp. 172-178, 2020. UCI(KEPA): I410-ECN-0101-2020-004-000454079.

[2] G. X. Wang and S. Y. Shin, “An improved text classification method for sentiment classification,” *Journal of Information and Communication Convergence Engineering*, vol. 17, no. 1, pp. 41-48, 2019. DOI: 10.6109/jicce.2019.17.1.41.

[3] X. F. Wang and H. C. Kim, “Text categorization with improved deep learning methods,” *Journal of Information and Communication Convergence Engineering*, vol. 16, no. 2, pp. 106-113, 2018. UCI(KEPA): I410-ECN-0101-2018-004-003109645.

[4] P. Patel and A. Thakkar, “The upsurge of deep learning for computer vision applications,” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 1, pp. 538-548, 2020. DOI: 10.11591/ijece.v10i1.pp538-548.

[5] M. A. Jishan, K. R. Mahmud, and A. K. A. Azad, “Natural language description of images using hybrid recurrent neural network,” *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, pp. 2932-2940, 2019. DOI: 10.11591/ijece.v9i4.pp2932-2940.

[6] H. S. Yi, K. H. N. Bui, and C. N. Seon, “A deep learning LSTM framework for urban traffic flow and fine dust prediction,” *Journal of Korean Institute of Information Scientists and Engineering*, vol. 47, no. 3, pp. 292-297, 2020. DOI: 10.5626/JOK.2020.47.3.292.

[7] B. C. Kim, S. H. Jung, M. S. Kim, J. G. Kim, H. S. Lee, and S. S. Kim, “Solar power generation forecasting based on LSTM considering weather conditions,” *Journal of Korean Institute of Intelligent Systems*, vol. 30, no. 1, pp. 7-12, 2020. DOI: 10.5391/JKIIS.2020.30.1.7.

[8] N. P. Shetty, B. Muniyal, A. Anand, S. Kumar, and S. Prabhu, “Predicting depression using deep learning and ensemble algorithms on raw twitter data,” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 3751-3756, 2020. DOI: 10.11591/ijece.v10i4.pp3751-3756.

Jongwon Lee

received the M.S. degree in 2016 and Ph. D. degree in 2019 from the Department of Computer Engineering of Pai Chai University, Korea. Since 2020, he has been with the National Science & Technology Information Service at the Korea Institute of Science and Technology Information, where he is currently a Postdoctoral Researcher. His current research interests include deep learning, machine learning, big data, and artificial intelligence.

Guanchen Wu

received his B.S and M.S. degrees in 2011 and 2014, respectively, from the Department of Control Engineering of Chongqing University of Posts and Telecommunications, China. From 2014 to 2019, he worked for Guizhou Communications Polytechnic as a teacher. Since 2019, he has been with the Department of Computer Engineering at Pai Chai University. His current research interests include Internet of Things, big data, and machine learning.

Hoekyung Jung

received his M.S. and Ph. D. degrees in 1987 and 1993, respectively, from the Department of Computer Engineering, Kwangwoon University, Korea. From 1994 to 1995, he worked for ETRI as a researcher. Since 1994, he has worked for the Department of Computer Engineering at Pai Chai University, where he is now a professor. His current research interests include machine learning, IoT, big data, and artificial intelligence.