JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 21, No. 11 : 2133~2138 Nov. 2017

XML 문서 키워드 가중치 분석 기반 문단 추출 모델

이종원[1] · 강인식[2] · 정회경[3*]

XML Document Keyword Weight Analysis based Paragraph Extraction Model

Jongwon Lee[1] · Inshik Kang[2] · Hoekyung Jung[3*]

[1]Department of Computer Engineering, Paichai University, Daejeon 35345, Korea
[2]Korea University of Media Arts, 312, Daehak-gil, Sejong, Korea
[3*]Department of Computer Engineering, Paichai University, Daejeon 35345, Korea

요 약

기존의 XML 문서나 다른 문서는 단어를 중심으로 분석이 진행되었다. 이는 형태소 분석기를 활용하여 구현이 가능하나 문서 내에 기재되어 있는 많은 단어를 분류할 뿐 문서의 핵심 내용을 파악하기에는 어려움이 있다. 사용자가 문서를 효율적으로 이해하기 위해서는 주요 단어가 포함되어 있는 문단을 추출하여 사용자에게 보여주어야 한다. 본 논문에서 제안하는 시스템은 정규화 된 XML 문서 내에 키워드를 검색하고 사용자가 입력한 키워드들이 포함되어 있는 문단을 추출하여 사용자에게 보여준다. 그리고 검색에 사용된 키워드들의 빈도수와 가중치를 사용자에게 알려주고 추출한 문단의 순서와 중복 제거 기능을 통해 사용자가 문서를 이해하는데 발생할 수 있는 오류를 최소화하였다. 제안하는 시스템은 사용자가 문서 전체를 읽지 않고 문서를 이해할 수 있게 하여 문서를 이해하는데 필요한 시간과 노력을 최소화할 수 있을 것으로 사료된다.

ABSTRACT

The analysis of existing XML documents and other documents was centered on words. It can be implemented using a morpheme analyzer, but it can classify many words in the document and cannot grasp the core contents of the document. In order for a user to efficiently understand a document, a paragraph containing a main word must be extracted and presented to the user. The proposed system retrieves keyword in the normalized XML document. Then, the user extracts the paragraphs containing the keyword inputted for searching and displays them to the user. In addition, the frequency and weight of the keyword used in the search are informed to the user, and the order of the extracted paragraphs and the redundancy elimination function are minimized so that the user can understand the document. The proposed system can minimize the time and effort required to understand the document by allowing the user to understand the document without reading the whole document.

키워드 : 문단 추출, 문서 분석, 압축, 키워드 가중치, 키워드 빈도수

Key word : Compression, Document Analysis, Keyword Frequency, Keyword Weight, Paragraph Extraction

Received 22 October 2017, **Revised** 29 October 2017, **Accepted** 04 November 2017

* **Corresponding Author** Hoekyung Jung(E-mail:hkjung@pcu.ac.kr, Tel:+82-42-520-5640)
Department of Computer Engineering, Paichai University, Daejeon 35345, Korea

Open Access https://doi.org/10.6109/jkiice.2017.21.11.2133 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

데이터의 양이 늘어남에 따라 문서 작성과 문서 이해에 대한 중요성이 증가하고 있다. 데이터가 급증함에 따라 새로운 기술 분야인 '빅 데이터'가 대두된 것처럼 작성된 문서를 이해하는데 도움을 주는 프로그램이 요구되고 있다.

문서의 종류는 많고 문서 작성의 목적도 다양하다. 반면에 문서를 이해하는데 시간이 필요한 것은 동일하다. 기존의 문서 분석 시스템은 형태소 분석기를 기반으로 개발된 시스템이 대부분이다. 이는 문서 작성에 사용된 단어들을 나열하고 사용 횟수를 보여줌으로써 사용자에게 해당 문서의 주요 단어를 알 수 있게 해준다[1-4]. 주요 단어를 알 수 있다고 해서 문서를 이해하는데 필요한 시간이 단축되는 것은 아니다. 이를 위해 다른 시스템들은 사용자가 입력한 검색어가 포함되어 있는 문서나 문단을 찾고 사용자에게 이를 보여준다[5-7]. 사용자가 문서를 이해하는데 도움이 되기는 하지만 문서를 이해하는데 필요한 시간을 줄이거나 효율성을 높이지는 못한다. 이를 위해서는 주요 단어를 형태소 분석기로 알아내는 것처럼 주요 문단을 추출하여 사용자에게 보여주는 기능이 필요하다.

본 논문에서 제안하는 시스템은 사용자가 문서를 이해하는데 필요한 시간을 줄이고 문서 이해에 대한 효율성을 높일 수 있게 한다. 문서의 전체 내용에서 문서 작성자가 기재한 키워드를 알려주고 사용자가 입력한 키워드가 포함되어 있는 문단을 추출한다. 키워드와 관련된 문단을 추출함으로써 중요도가 낮은 문단을 배제하고 이로 인해 사용자는 문서를 이해하는데 필요한 시간이 단축시킬 수 있다.

Ⅱ. 시스템 설계

본 장에서는 제안하는 시스템의 설계를 다룬다. 시스템은 사용자가 검색을 원하는 XML 문서를 입력하면 해당 문서를 읽어온 뒤 키워드를 입력하면 키워드가 포함되어 있는 문단을 추출한다. 문단을 추출하고 중복되는 문단이 있을 경우 제거한다. 그리고 XML 문서 내에서 문단의 순서를 유지한 채 결과를 도출해낸 뒤 사용자에게 보여준다. 이러한 기능들을 구현하기 위해 그림

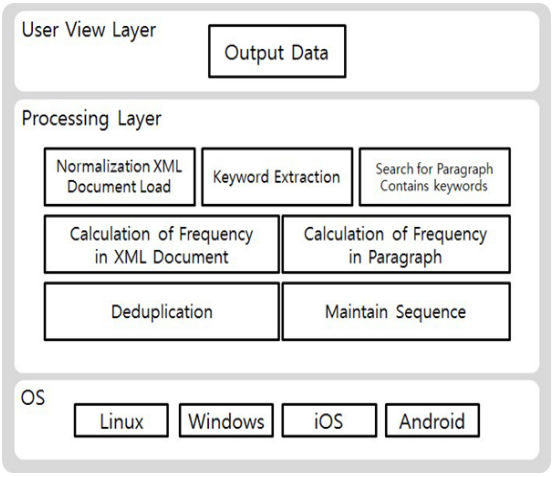

Fig. 1 System Architecture

1과 같이 시스템의 구조를 설계하였다.

프로그램이 필요로 하는 기능들은 사용자가 입력한 XML 문서를 불러오는 기능과 키워드를 추출하는 기능, 입력한 키워드를 포함하고 있는 문단을 추출하는 기능, 문단의 순서를 유지하는 기능, 문단의 중복을 확인하고 제거하는 기능이 필요하다. 이를 위해 시스템은 3개의 계층 구조로 설계하였다. 제안하는 시스템은 Java로 구현하기 때문에 OS에 종속되지 않고 다양한 환경에서 실행이 가능하다.

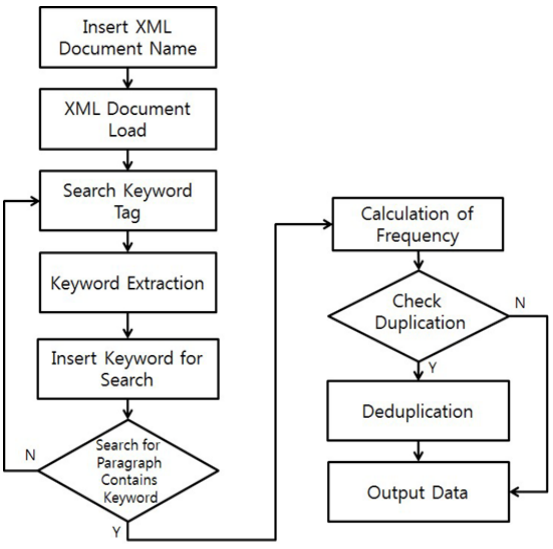

Fig. 2 System Flowchart

그림 2는 시스템의 흐름이다. 시스템은 사용자가 XML 문서의 파일명을 입력하면서 시작된다. 해당 파일명의 XML 문서를 읽어오고 해당 문서 내에 기입 되어 있는 키워드 태그를 찾아서 키워드들을 저장한다. 그리고 사용자에게 키워드들을 보여줌으로써 사용자는 검색을 원하는 키워드의 개수와 키워드를 입력하게 된다. 사용자가 키워드 입력을 마치면 해당 키워드들이 포함되어 있는 문단을 검색하여 추출한다. 문단을 추출한 뒤에는 검색한 키워드들의 빈도수들을 계산하여 출력한다. 키워드의 빈도수는 2가지 종류로 출력하는데 추출한 문단 전체에서의 빈도수와 특정 문단에서의 빈도수이다. 문단 전체에서의 빈도수는 키워드들 간의 가중치를 비교하기 위해서 횟수와 퍼센트로 표현하고 문단 내에서의 빈도수는 특정 문단에 어떤 키워드가 사용되었는지를 명시하기 위해 횟수로 표현한다.

Ⅲ. 시스템 구현

본 장에서는 제안하는 키워드 가중치 기반 XML 문서 문단 추출 시스템의 구현을 다룬다. 또한 효율성을 검증한다. 시스템이 시작되면 XML 문서의 파일명을 입력 받게 된다. 그림 3은 해당 기능의 흐름을 나타낸 것이다.

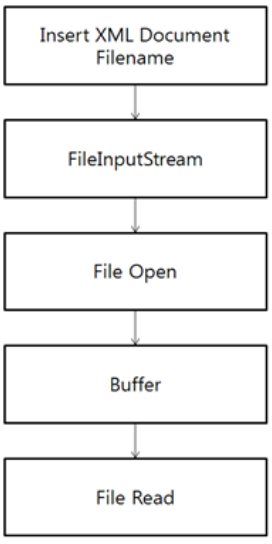

Fig. 3 XML Document File Open Flowchart

사용자가 XML 문서의 파일명을 입력하면 Java FileInputStream 클래스의 기능으로 파일을 열고 Java Buffer를 활용하여 파일을 읽기 시작한다. 그림 4는 키워드 추출의 흐름을 나타낸 것이다.

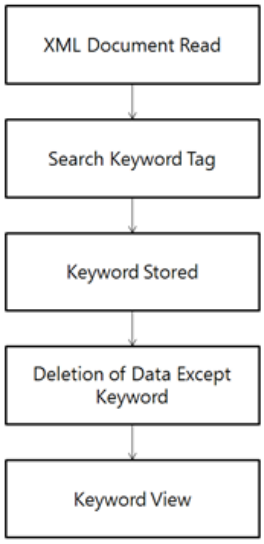

Fig. 4 Keyword Extraction Flowchart

키워드 추출이 완료된 뒤 검색을 원하는 키워드의 개수를 입력 받게 된다. 해당 XML 문서의 키워드를 보여주고 사용자는 키워드들 중에서 검색을 원하는 키워드를 입력하게 된다. 그림 5는 해당 기능의 흐름이다.

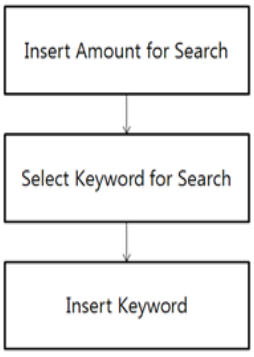

Fig. 5 Insert Keyword Flowchart

검색을 위한 키워드 입력이 완료되면 시스템은 키워드가 포함되어 있는 문단들을 검색하고 추출한다. 그림

6은 키워드 빈도수로 비교하는 기능의 흐름을 나타낸 것이다.

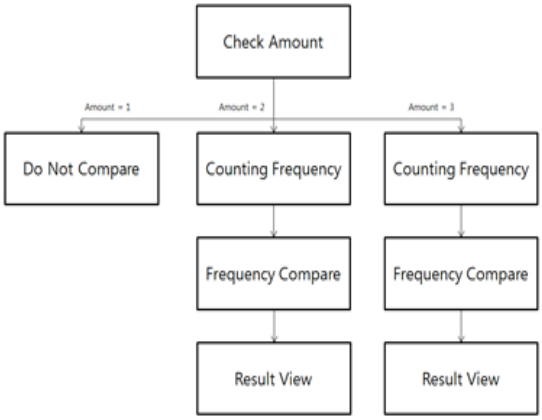

Fig. 6 Frequency Compare Flowchart

키워드들의 비교 내용 출력 다음에는 키워드 빈도수와 가중치를 보여주고 중복 제거된 문단의 수를 알려준다. 그리고 각 키워드들의 빈도수와 빈도수를 퍼센트로 환산하여 가중치를 보여준다. 그림 7은 중복 제거 기능의 흐름을 나타낸 것이고 그림 8은 사용자가 입력한 키워드들의 검색 결과를 보여준다.

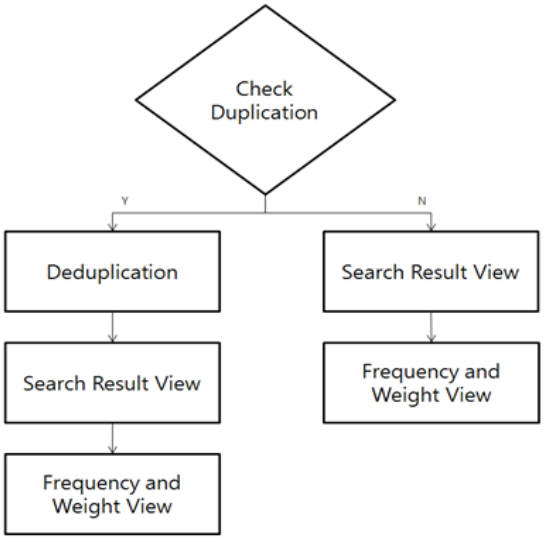

Fig. 7 Frequency Compare Flowchart

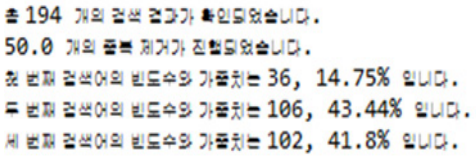

Fig. 8 Screen of Compare Output

시스템이 사용자에게 검색 결과를 보여준 뒤 문단 내에 키워드들의 빈도수와 추출된 문단을 보여준다. 시스템은 사용자에게 기존 시스템에 비해 압축률이 높고 검색한 키워드와 관련된 문단을 보여줌으로써 정확도가 높은 서비스를 제공하게 된다. 그림 9는 문단 내에 키워드들의 빈도수와 추출된 문단을 출력하는 기능의 흐름을 나타낸 것이고 그림 10은 해당 기능의 결과이다.

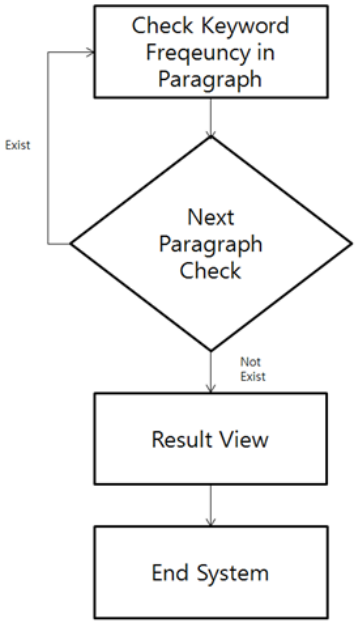

Fig. 9 Check Keyword Frequency in Paragraph Flowchart

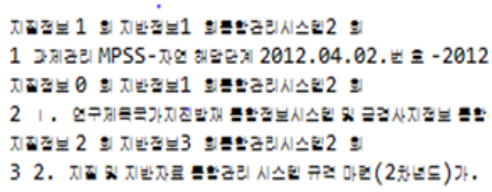

Fig. 10 Screen of Output

Ⅳ. 고 찰

기존의 문서 분석 시스템들은 문서 작성에 사용된 단어들을 분류하고 빈도수를 확인하는 동작을 주목적으로 하였다. 또한 사용자가 입력한 검색어가 포함되어 있는지를 확인하는 것이 대부분이었다. 이는 문서를 이해하는데 걸리는 시간을 줄이지 못하는 문제점이 있었다. 이를 해결하기 위해 본 논문에서는 사용자가 입력한 키워드가 포함되어 있는 문단들을 추출하고 키워드들의 빈도수와 가중치를 계산하고 사용자에게 보여준다. 또한 추출된 문단들을 중복 제거 처리를 진행하고 사용자에게 추출한 문단들을 보여줌으로써 사용자가 검색한 키워드 외에 내용을 배제한 체 문서를 읽을 수 있게 하였다. 그림 11은 제안하는 시스템의 효율성을 검증하기 위해 기존 시스템과의 비교 실험을 진행한 뒤 결과를 정리한 것이다.

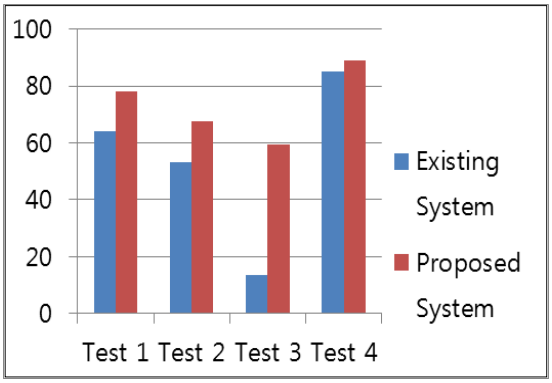

Fig. 11 Test Result Graph

실험은 4개의 정규화 된 XML 문서로 진행하였다. 첫 번째 실험에서 제안하는 시스템은 압축률이 약 78.0 퍼센트였고 기존 시스템보다 약 13.9퍼센트 높은 압축률을 보였다. 두 번째 실험에서 제안하는 시스템은 압축률이 약 67.4 퍼센트였고 기존 시스템보다 약 14.0퍼센트 높은 압축률을 보였다. 세 번째 실험에서 제안하는 시스템은 압축률이 약 59.4 퍼센트였고 기존 시스템보다 약 45.9퍼센트 높은 압축률을 보였다. 네 번째 실험에서 제안하는 시스템은 압축률이 약 88.8 퍼센트였고 기존 시스템보다 약 3.9퍼센트 높은 압축률을 보였다.

교집합 연산처리를 진행하는 기존 시스템은 추출 문단의 개수가 적지만 정확도가 낮아서 정확한 문단 추출이 어려운 문제점이 있었다. 이를 해결하기 위해 대부분의 문서 분석 시스템들은 합집합 연산처리를 통해 문서 분석을 진행한다. 그러나 합집합 연산 처리를 하는 기존 시스템들은 압축률이 낮아 사용자가 읽어야 할 문단이 많은 단점이 있었다. 제안하는 시스템은 합집합 연산처리로 문단을 검색하고 교집합 연산처리로 중복되는 문단을 제거하였고 사용자가 입력한 키워드의 빈도수와 가중치를 보여줌으로써 사용자의 문서 이해를 효율적으로 진행할 수 있게 하였다.

Ⅴ. 결 론

본 연구에서 제안하는 시스템은 사용자가 검색을 원하는 XML 문서를 입력하면 해당 문서를 불러온 뒤 키워드들을 보여준다. 그리고 사용자가 검색을 원하는 키워드와 키워드 개수를 입력하면 이에 따라 키워드가 포함되어 있는 문단들을 추출한다. 문단을 추출한 뒤 시스템은 각 키워드들의 빈도수와 가중치를 계산하여 해당 검색에서 사용된 키워드들의 가중치를 보여준다. 또한 검색에 사용된 키워드들이 포함되어 있는 문단의 개수를 알 수 있고 문단의 순서 유지 기능과 중복 제거 기능을 통해 XML 문서의 내용이 훼손되지 않은 상태로 결과를 볼 수 있다. 이로 인해 기존 시스템에 비해 정확도와 압축률이 높기 때문에 XML 문서로 보고서나 논문 등의 자료를 분석하는 연구 분야에서 사용자들에게 문서를 이해하는데 높은 효율성을 보일 것으로 예상된다. 또한 가중치와 빈도수 비교, 중복 제거 처리, 순서 유지 기능 등 문서를 이해하는데 도움이 되는 기능들을 제공함으로써 문서 작성 및 분석과 관련된 분야에 높은 파급 효과를 제공할 것으로 사료된다.

향후 연구로는 다양한 실험을 통해 시스템의 편리성과 효율성을 검증하고 UI 수정을 진행하여야 한다.

ACKNOWLEDGMENTS

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

REFERENCES

[1] B. J. Noh, Z. S. Xu, J. G. Lee, D. H. Park, Y. H. Chung, "Keyword Network Based Repercussion Effect Analysis of Foot-and-Mouth Disease Using Online News," *Korean Institute of Information Technology*, vol. 14, no. 9, pp. 143-152, Sep. 2016.

[2] S. J. Choi, J. W. Lee, "A Morphological Analysis Method of Prediction place-Event Performance by Online News Titles," *Korea Association of Community Welfare Studies*, vol. 21, no. 1, pp. 15-32, Feb. 2016.

[3] H. S. Ha, B. Y. Hwang, "Keyword Filtering about Disaster and the Method of Detecting Area in Detecting Real-Time Event Using Twitter," *Korea Information Processing Society*, vol. 5, no. 7, pp. 345-350, Jul. 2016.

[4] J. C. Shin, C. Y. Ock, "A Korean Morphological Analyzer using a Pre-analyzed Partial Word-phrase Dictionary," *Korean Institute of Information Scientists and Engineering*, vol. 39, no. 5, pp. 415-424, May 2012.

[5] S. H. Na, J. I. Kim, E. J. Lee, P. K. Kim, "A Study on the Short Text Categorization using SNS Feature Informations," *Korean Institute of Information Technology*, vol. 14, no. 6, pp. 159-165, Jun. 2016.

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

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

이종원(Jongwon Lee)

2014년 배재대학교 컴퓨터공학과(공학사)
2016년 배재대학교 컴퓨터공학과(공학석사)
2016년 ~ 현재 배재대학교 컴퓨터공학과(박사과정)
※관심분야 : U-Healthcare, 빅 데이터, IoT

강인식(Inshik Kang)

1992년 청주대학교 연극영화학과(문학사)
2016년 ~ 현재 배재대학교 컴퓨터공학과 석사과정
1992년 ~ 1997년 MBC 문화방송
1997년 ~ 2005년 ITV 경인방송
2006년 MBC 미디어텍
2008년 ~ 현재 휴미디어 대표
2013년 ~ 현재 한국영상대학교 교수
2013년 ~ 현재 한국영상대학교 촬영조명과 교수
※관심분야 : 영상신호, 오디오신호, 방송국제신호제작(IS)

정회경(Hoekyung Jung)

1985년 광운대학교 컴퓨터공학과(공학사)
1987년 광운대학교 컴퓨터공학과(공학석사)
1993년 광운대학교 컴퓨터공학과(공학박사)
1994년 ~ 현재 배재대학교 컴퓨터공학과 교수
※관심분야 : 멀티미디어정보처리, XML, Semantic Web, Ubiquitous Computing, USN, IoT