문단 가중치 분석 기반 본문 영역 선정 알고리즘

이종원 · 유성종 · 김도안 · 정회경

배재대학교

Keyword Weight based Paragraph Extraction Algorithm

Jongwon Lee · Seongjong Yu · Doan Kim · Hoekyung Jung

PaiChai University

E-mail : {starjwon, noelrika, doan0105}@naver.com, hkjung@pcu.ac.kr

요 약

기존의 문서 분석 시스템들은 형태소 분석기나 TF-IDF 기법을 통해 단어 위주의 분석을 진행하였다. 이러한 시스템들은 키워드들의 가중치를 계산하여 주요 키워드를 도출할 수 있는 장점이 있다. 이에 반해 문서의 내용을 분석하기에는 구조적인 한계로 인해 부적합한 실정이다. 이를 해결하기 위해 본 논문에서 제안하는 알고리즘은 문서 내에 있는 문단들의 가중치를 계산한 뒤 문단들을 영역별로 분할한다. 그리고 분할된 영역별로 중요도를 계산하여 해당 문서 내에 가장 중요한 문단들이 있는 영역을 사용자에게 알려준다. 이를 통해 사용자는 기존의 문서 분석 시스템들을 사용할 때보다 문서를 분석하기에 적합한 서비스를 제공받을 것으로 사료된다.

ABSTRACT

Traditional document analysis systems used word-based analysis using a morphological analyzer or TF-IDF technique. These systems have the advantage of being able to derive key keywords by calculating the weights of the keywords. On the other hand, it is not appropriate to analyze the contents of documents due to the structural limitations. To solve this problem, the proposed algorithm calculates the weights of the documents in the document and divides the paragraphs into areas. And we calculate the importance of the divided regions and let the user know the area with the most important paragraphs in the document. So, it is expected that the user will be provided with a service suitable for analyzing documents rather than using existing document analysis systems.

키워드

Document Analysis, Keyword Weight, Paragraph Centrality, Search

Ⅰ. 서 론

문서를 분석하여 새로운 문서를 작성하거나 가치를 창출하는 연구가 진행되고 있다. 형태소 분석기를 통해 문서 내에 있는 단어들을 나열하거나 텍스트 마이닝을 통해 단어와 단어, 문장과 문장, 문단과 문단을 분석하여 새로운 의미를 부여하는 연구가 대표적이다. 이에 반해 문서를 이해하는데 도움을 주는 시스템은 연구 중인 실정이다. 기존의 문서 분석 시스템들은 형태소 분석기를 기반으로 단어 중심의 분석을 진행하기 때문에 문서의 내용을 사용자가 이해하는 데 도움을 주기 어려운 문제점이 있었다[1,2]. 다른 시스템들은 검색어가 포함되어 있는 문서나 문단을 찾은 뒤 사용자에게 보여준다[3]. 이러한 시스템들도 사용자가 문서를 이해하는데 도움을 주기는 어렵다.

이러한 문제점을 해결하기 위해 본 논문에서 제안하는 시스템은 문서의 키워드를 추출한 뒤 키워드가 포함되어 있지 않은 문단들을 제거한다. 그리고 키워드의 가중치를 비교하여 가중치가 낮은 키워드만을 포함한 문단을 제거한 뒤 문단들을 10개의 영역으로 분할하고 분석하여 핵심 영역을 사용자에게 보여준다. 이로 인해 사용자는 기존 시스템에 비해 문서 이해에 필요한 시간을 단축시킬 수 있을 것으로 사료된다.

Ⅱ. 시스템 설계

본 장에서는 제안하는 시스템의 설계를 기술한다. 그림 1은 시스템의 구조를 나타낸다.

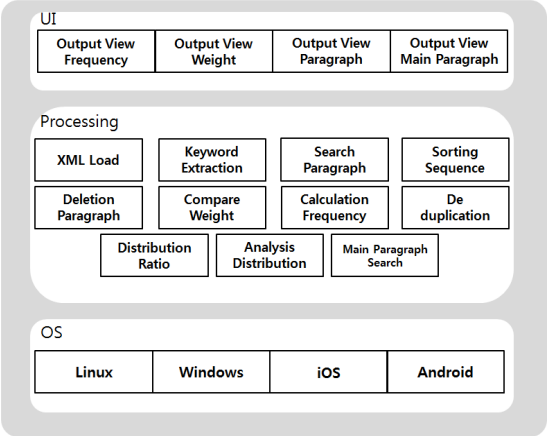

그림 1. 시스템 구조도

그림 1은 시스템의 구조이다. 시스템이 필요로 하는 기능들은 다음과 같다. 분석을 위해 사용자가 입력한 XML 문서를 불러오는 기능과 해당 문서의 키워드를 추출하여 사용자에게 보여주는 기능, 사용자가 입력한 키워드를 포함하고 있는 문단들을 추출하는 기능, 문단들의 순서를 유지하는 기능, 문단들의 중복을 확인하고 이를 제거하는 기능, 문단들을 10개의 영역으로 나누는 기능, 영역 내에 있는 키워드들의 빈도수를 계산하는 기능, 핵심 영역을 찾는 기능 등이 필요하다.

시스템이 시작되면 사용자는 XML 문서의 파일명을 입력한다. 시스템은 사용자가 입력한 파일명의 XML 파일을 불러온 뒤 키워드를 추출하여 사용자에게 보여준다. 사용자는 추출된 키워드들 중 3개의 키워드를 입력하여 검색을 시작한다. 시스템은 3개의 키워드들이 포함되어 있는 문단들을 추출한 뒤 문단 순서 유지, 중복 제거, 가중치가 낮은 키워드를 포함하고 있는 문단들을 제거하여 압축률을 높인다. 그리고 문단들의 영역을 나눈 뒤 핵심 영역을 도출하여 사용자에게 보여준다.

제안하는 시스템은 3개의 계층 구조로 설계하였다. Java로 구현하기 때문에 OS에 종속되지 않고 다양한 환경에서 실행이 가능하도록 하였다.

III. 결 론

본 논문에서 제안하는 시스템은 사용자가 분석을 원하는 XML 문서를 입력하면 해당 문서를 불러온 뒤 문서의 키워드들을 사용자에게 보여준다. 그리고 사용자가 검색을 원하는 키워드를 입력하면 키워드가 포함되어 있는 문단들을 추출한다. 문단들을 추출한 뒤 시스템은 각 키워드들의 빈도수와 가중치를 계산하고 가중치가 낮은 키워드만을 포함하고 있는 문단들을 제거하여 압축률을 높인다. 또한 문단들을 10개의 영역으로 분할하여 영역들의 중요도를 비교한다. 그리고 가장 높은 중요도를 나타낸 영역을 해당 문서의 핵심 영역으로 선정한 뒤 사용자에게 알려주는 기능을 수행한다. 이로 인해 XML 문서 형태의 보고서나 논문 등의 자료를 분석하고 활용하는 연구 분야에서 문서를 이해하는데 도움을 줄 수 있을 것으로 예상된다.

참고문헌

[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. 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.

[3] S. D. Kim, S. B. Yang, “A Comparative Analysis between General Comments and Social Comments on an Online News Site.” *The Korea Contents Association*, vol. 15, no. 4, Apr. 2015.