효율적인 문서 구성을 위한 TF-IDF 알고리즘 기반 문서 제안 시스템의 설계

김영훈[O], 박승민[*], 조대수[*]
[O]동서대학교 소프트웨어학과,
[*]동서대학교 소프트웨어학과
e-mail: dscho@dongseo.ac.kr

Design of Document Suggestion System based on TF-IDF Algorithm for Efficient Organization of Documentation

Young-Hoon Kim[O], Seung-Min Park[*], Dae-Soo Cho[*]
[O]Dept. of Software, Dongseo University,
[*]Dept. of Software, Dongseo University

● 요 약 ●

빠르게 변하는 환경에 맞춰 평생 교육이 일반화되고 개인에게 요구되는 학습량은 많아지고 있으며 높아진 학습량에 맞게 학습 시간 단축과 효율적인 학습을 위한 학습 방법을 선택하는 것이 중요해지고 있다. 본 논문에서는 학습 정리를 위해 작성한 문서를 분석하여 해당 문서와 관련된 문서를 제안하고 본 문서와 엮어 학습을 위한 문서 묶음을 만들 수 있는 시스템을 제안한다. 문서의 유사도, 중요도를 구할 수 있는 TF-IDF를 이용하여 문서를 분석해 키워드를 추출한 다음 그와 관련된 문서를 제안하고 문서 묶음을 만들어 조회할 수 있도록 한다. 이 시스템은 학습 정리 시 관련 문서를 함께 볼 수 있도록 하고, 필요하다면 묶음으로 만들어 효과적인 학습을 위한 도구로 이용할 수 있다.

키워드: TF-IDF(Term Frequency-Inverse Document Frequency), 문서 분석(Document Analysis), 유사도(Similarity), 키워드 추출(Keyword Extraction), 텍스트 마이닝(Text Mining)

I. Introduction

모든 것이 빠르게 변하는 현대의 환경에 적응하기 위해 평생 교육이라는 단어가 일반화되고[1] 개개인에게 요구되는 학습의 양은 많아지고 있다. 높아진 학습량에 따라 학습 시간을 단축하고 학습 효율을 높일 수 있는 학습 방법을 선택하는 것이 중요해지고 있다. 본 논문은 학습 정리를 위해 작성한 문서를 분석하여 해당 문서와 관련된 문서를 제안하고 본 문서와 엮어 학습을 위한 문서 묶음을 만들 수 있는 시스템에 관한 내용을 기술한다. 이 시스템은 작성된 문서의 단어 빈도수를 파악하여 그와 비슷하거나 관련 있는 다른 문서를 제안하고 문서 묶음을 만들 수 있도록 한다.

II. Preliminaries

문서를 자동 분류하는 알고리즘은 문서 내부의 단어로부터 문서 벡터를 만들어 학습에 이용하고 문서에 범주를 지정한다. TF-IDF는 단어별 가중치로 문서의 특징을 표현하여 두 문서 간 유사도를 비교하거나 문서의 핵심어를 추출하는 방법이다[2][3].

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (1)$$

식 (1)에서 TF(d,t)는 문서 d에서 단어 t의 등장 횟수를 말하며 출현 횟수가 많을수록 중요한 단어로 판단한다. IDF(t,D)는 다른 문서에서 흔하게 나타나는 단어일수록 가중치를 낮추기 위한 값이며 전체 문서 D의 수를 단어 t가 포함된 문서의 수 + 1의 값으로 나눈 다음 로그를 취하는 것이 일반적이다. 분모에 1을 더하는 이유는 t가 포함된 문서의 수가 0일 때 분모가 0이 되는 것을 방지하기 위함이며 TF-IDF는 TF(d,t)와 IDF(t,D)의 곱으로 나타낸다.

III. The Proposed Scheme

본 논문에서 제안하는 TF-IDF 알고리즘 기반 문서 제안 시스템은 사용자가 작성한 문서를 분석하여 키워드를 추출하고 키워드와 관련 있는 문서를 제안하여 문서 묶음을 만들 수 있도록 Fig 1.과 같은 절차를 따른다.

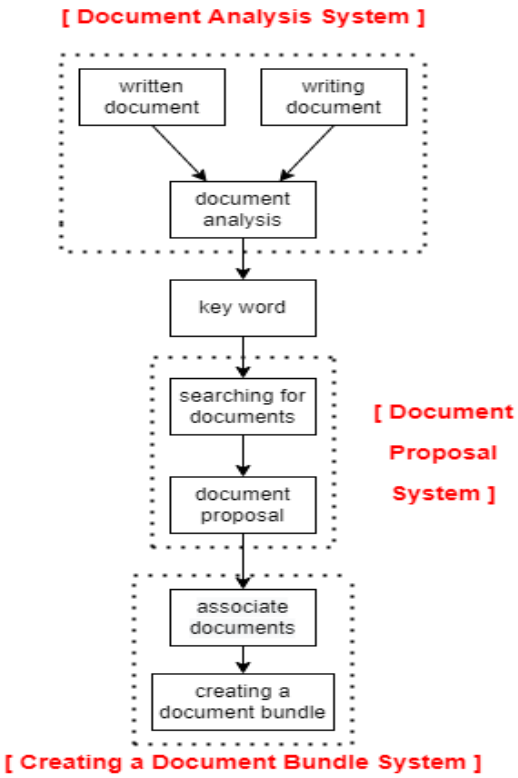

Fig. 1. System Structure

Document Analysis System은 사용자가 문서를 작성 중일 때 작성한 단어 빈도수를 파악하여 실시간으로 키워드를 추출할 수 있으며 작성이 완전히 끝난 문서에 대해서는 사용자가 현재까지 작성한 문서가 하나일 경우 해당 문서 내부의 단어 빈도수를 측정해 중요도를 결정하고 둘 이상일 경우 TF-IDF 알고리즘을 적용하여 특정 단어가 다른 문서에서 얼마나 흔하게 나타나는지를 가중치에 반영하여 중요도를 결정한 다음 키워드를 추출할 수 있다. 문서를 대표하는 키워드가 결정되면 Document Proposal System은 각 문서의 키워드를 이용해 미리 설정된 목표 사이트에서 관련 문서를 찾아 사용자가 참고하거나 문서 묶음으로 만들 수 있도록 제안한다. 사용자는 이렇게 제안된 문서를 단순히 열람하거나 선택적으로 Creating a Document Bundle System을 이용하여 작성한 문서와 연관되는 문서 묶음으로 만들어 추후 본 문서를 열람할 때 참고할 수 있도록 만들 수 있다.

작성 중이거나 작성이 끝난 문서로부터 추출되는 키워드를 이용한 관련 문서 제안을 통해 사용자는 문서를 작성할 때 관련된 문서를 참고하여 문서 작성에 활용하거나, 작성이 끝난 문서와 관련된 문서를 참고 자료로 나타낼 수 있어 문서 키워드와 관련된 내용에 대해 효율적인 학습이 가능할 것이다.

IV. Conclusions

평생 교육의 시대에서 학습 효율을 높일 수 있는 학습 방법을 선택하는 것은 중요하다. 본 논문에서는 학습 정리를 위해 작성한 문서 분석 후 관련 문서를 제안 및 본 문서와 엮어 문서 묶음을 만들 수 있는 시스템을 제안했다. 제안된 시스템을 통해 학습 정리를 위한 문서를 작성할 때 제안된 관련 문서들을 학습 참고 자료로 활용할 수 있으며 묶음으로 만들어 효과적인 학습을 위한 도구로 이용할 수 있을 것이다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음(2019-0-01817)

REFERENCES

[1] Laal, Marjan, Peyman Salamati. “Lifelong learning; why do we need it?” Procedia-Social and Behavioral Sciences, Vol. 31, pp.399-403, 2012.

[2] You Eun-Soon, Choi Gun-Hee, Kim Seung-Hoon. “Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels” Journal of the Korea Society of Computer and Information, Vol. 20, pp.121-129, 2015.

[3] QAISER Shahzad, ALI Ramsha. “Text mining: use of TF-IDF to examine the relevance of words to documents.” International Journal of Computer Applications, Vol. 181, pp.25-29, 2018.