학습 문서에 최적화된 자동문서 분류 시스템 설계 및 구현

*강미영, **강대욱
*전남대학교 소프트웨어공학 협동과정 **전남대학교 전산학과
*juriaya@chonnam.ac.kr , **dwkang@chonnam.ac.kr

Automatic Document Classification System for Studying Document

*Mi-Young Kang, **Dae-Wook Kang
*Dept. of Software Engineering Cooperation Course. Chonnam Univ
**Dept. of Computer Science, Chonnam Univ

요 약

자동문서 분류란 문서의 내용에 기반을 두어 미리 정의되어 있는 범주에 문서를 자동으로 할당하는 작업이다. 이를 위해 형태소 분석을 이용한 단어빈도, 구문분석, 의미분석 등을 이용한 다양한 기법들이 제시되어 왔으며 이를 적용하여 사용자와 관리자 모두의 편의성을 높일 수 있다. 하지만 주제에 대부분의 키워드가 포함된 학습문서의 특성을 무시하고 기존의 방법을 그대로 적용하기엔 시스템의 부하와 실행시간 증가와 같은 기존의 문제점을 답습하게 된다. 본 논문에서는 학습문서의 특성을 분석하고, 존재하는 학습문서 등록시스템에 학습문서의 특성에 적합한 자동문서 분류 기법을 제안한다. 제안한 기법을 적용한 학습문서 등록 시스템은 문서분류의 정확성을 높이고 사용자의 편의성을 증대시킬 수 있다.

I. 서 론

온라인상에서 얻을 수 있는 정보의 양이 늘어남에 따라 효율적인 정보의 관리와 검색이 요구되고 있다. 컴퓨터를 통해 접하게 되는 웹 문서 등의 전자문서를 분류하는 것은 쉽지 않은 일이며, 필요한 정보를 효율적으로 습득하기 위한 정보접근을 위해 문서에 대한 분류 작업이 필요하다. 문서 분류를 수작업으로 하는 경우 사람의 노력, 시간, 비용 면에서 비효율적이며, 분류하는 사람의 지식과 수준이 다르기 때문에 같은 문서를 다르게 분류 할 수 있다. 이렇게 잘못 분류된 문서는 사용자에게 혼란을 야기 시키므로 자동문서 분류 시스템의 필요성이 대두되고 있다.

문서 분류(document classification) 혹은 텍스트범주화(Text categorization)는 미리 정의해둔 일정한 수의 카테고리들에 문서들을 분류하는 작업을 말한다. 분류 작업을 자동으로 하기 위해서는 자연어 이해 및 처리기술이 필수적이다. 그러나 현재의 자연어 처리 기술로는 만족할만한 분류 결과를 얻기 어는 어려운 실정이다. 자동문서 분류는 이러한 작업들을 수작업이 아닌 컴퓨터를 이용하여 자동으로 문서를 분류하는 것을 말한다.

자동문서 분류 시스템에서 가장 보편적인 방법이 문장의 구분 없이 전체 문서에 출현한 각 자질의 빈도(TF: Term Frequency)와 역문헌빈도(IDF: Inverse Document Frequency) 등을 이용하여 표현하는 방법이다.

본 논문에서는 웹상의 여러 문서 중에서 교육용 학습 문서의 특징을 알아보고 TF, IDF보다 발전된 방향으로 학습 문서를 자동 분류하는 방법을 제안한다.

2장에서는 문서 분류 분야에서 수행되었던 연구들에 대해 살펴보고,