문서 범주화 효율성 제고를 위한 정보원 평가에 관한 연구*

A Study on Information Resource Evaluation for Text Categorization

정은경 (EunKyung Chung)**

초록

이 연구는 색인가가 주제 색인하는 과정에서 참조하는 여러 문서구성요소를 문서 범주화의 정보원으로 인식하여 이들이 문서 범주화 성능에 미치는 영향을 살펴보는데 그 목적이 있다. 이는 기존의 문서 범주화 연구가 전문(full text)에 치중하는 것과는 달리 문서구성요소로서 정보원의 영향을 평가하여 문서 범주화에 효율적으로 사용될 수 있는지를 파악하고자 한다. 전형적인 과학기술 분야의 저널 및 회의록 논문을 데이터 집합으로 하였을 때 정보원은 본문정보 중심과 문서구성요소 중심으로 나뉘어 질 수 있다. 본문정보 중심은 본론 자체와 서론과 결론으로 구성되며, 문서구성요소 중심은 제목, 인용, 출처, 초록, 키워드로 파악된다. 실험 결과를 살펴보면, 인용, 출처, 제목 정보원은 본문 정보원과 비교하여 유의한 차이를 보이지 않으며, 키워드 정보원은 본문 정보원과 비교하여 유의한 차이를 보인다. 이러한 결과는 색인가가 참고하는 문서구성요소로서의 정보원이 문서 범주화에 본문을 대신하여 효율적으로 활용될 수 있음을 보여주고 있다.

Abstract

The purpose of this study is to examine whether the information resources referenced by human indexers during indexing process are effective on Text Categorization. More specifically, information resources from bibliographic information as well as full text information were explored in the context of a typical scientific journal article data set. The experiment results pointed out that information resources such as citation, source title, and title were not significantly different with full text. Whereas keyword was found to be significantly different with full text. The findings of this study identify that information resources referenced by human indexers can be considered good candidates for text categorization for automatic subject term assignment.

키워드: 문서범주화, 자동색인, 정보원, Text Categorization, 주제색인과정

---

* 본 고는 미국 노스텍사스대학교 박사 논문 일부를 요약한 것임

** 이화여자대학교 사회과학대학 문헌정보학 전임강사 (echung@ewha.ac.kr)

1. 서 론

전통적으로 주제어는 중요한 정보 접근 점으로 사용되어 왔다. 정보접근을 위한 주제어는 일반적으로 두 가지 방식으로 생성될 수 있다. 첫째, 색인가에 의해서 할당된 주제 색인어와 둘째, 자동으로 본문에서 추출된 키워드 색인어로 구분된다. 이 중에서 색인가에 의해 할당되는 주제 색인어는 사용자에게 보다 의미 있는 접근 점을 제공하고 동일한 주제 분야의 문서를 한 곳에 모은 데 (Collocate) 있어서 중요한 역할을 한다(O'Connor, 1996). 주제 색인어 생성은 일반적으로 전문 색인가의 주제 분석과 주제어 선정 과정을 통해 이루어진다. 이러한 과정은 색인가의 전문 분야 지식과 해당 통제언어 구조에 대한 이해를 필요로 하기 때문에 비용과 시간 면에서 대규모 전문 데이터베이스 서비스에서 주로 이용되고 있는 현실이다. 따라서 색인가에 의하여 생성되는 주제어는 고비용과 제한적인 처리 속도가 한계점으로 지적되어 왔다. 더욱이 최근의 인터넷과 디지털 기기의 발전으로 정보의 급속한 팽창이 가속되어 왔으며, 이에 비례하여 지속적인 주제 색인에 대한 요구로 인해 자동 색인어 할당에 관한 연구가 활발히 진행되고 있다.

특히 기계 학습(Machine Learning) 기반 컴퓨터 알고리즘을 이용하여 자동으로 색인어를 부여하는 문서 범주화 (Text Categorization)기술이 비용과 시간상의 제한점을 극복하기 위해서 도입되어 연구 및 적용되어 왔다. 문서 범주화는 컴퓨터 학습 알고리즘이 기존의 문서와 이에 할당된 주제어 집합을 이용하여 학습한 후 자동으로 새로운 문서에 최적의 주제어를 부여하는 방식이다. 이러한 문서 범주화 연구는 일반적으로 본문(full text)을 이용하여 효율적인 알고리즘의 개발, 성능 향상을 위한 자질 선정 (feature selection), 및 특정 문헌집단 최적화 등에 집중되어 왔다 (Cunningham, Witten, & Litten, 1999; Sebastiani, 2002; 2005). 이러한 연구의 집중은 색인가가 수행하는 색인과정을 지나치게 단순하게 인식하는데서 기인한다고 볼 수 있다(Moens, 2000). 이 연구는 기존의 전문 중심의 문서 범주화 연구와 달리 실제로 색인가가 거치는 색인과정에 관한 고찰을 포함하고 있다. 특히, 색인가가 색인과정에서 참조하는 문서의 구성 요소를 문서 범주화의 중요한 정보원(Information Resources)으로 인식하여 궁극적으로 문서 범주화 효율성을 제고하고자 한다. 이 연구는 색인과정에 관한 고찰과 인식으로 기존의 문서 범주화 연구에서 다루지 않았던 새로운 정보원을 발굴하여 문서 범주화 성능과 효율을 향상시킬

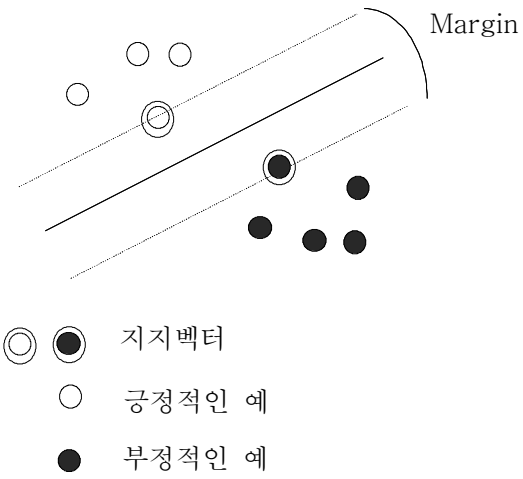
수 있는 도구로 사용될 수 있음을 제시하고자 한다.

2.문서범주화

2.1 지지벡터기계 (SVM)

기계학습(Machine Learning) 기반의 문서 범주화는 기존의 문서와 이에 할당된 주제어를 기계학습 알고리즘을 통해 파악한 후 새로운 문서에 학습된 알고리즘(classifier)을 통해 최적의 주제어를 자동으로 부여하는 기법이다. 즉, 문서 범주화는 할당된 주제어와 해당 문서와의 관계를 활용하여 새로운 문서에 자동으로 주제어를 할당하는 과정을 의미한다. 주로 사용되는 문서 범주화 알고리즘은 나이브 베이즈(Naive Bayes), 지지벡터기계(Support Vector Machine), 뉴럴 네트웍(Neural Network), kNN (k-nearest neighbors)등이 있다 (Sebastini, 2002).

다양한 문서 범주화 알고리즘 중에서 지지벡터기계 (SVM) 알고리즘은 Joachims (1998)에 의해 소개되었다. 지지벡터기계는 안정성과 신뢰할 만한 성능으로 인해 여러 분야의 문서 범주화 연구와 개발에 사용되어 왔다 (Sebastiani, 2005). <그림 1>에서 살펴볼 수 있는 바와 같이, 지지벡터기계 알고리즘은 궁극적으로 부정적인 예와 긍정적인 예를 찾아 양방간의 최대 간격(Margin)을 이루는 지지벡터를 찾아내는 것이다. 즉, 최대 간격을 찾아낼수록 부정적인 지지벡터와 긍정적인 지지벡터를 구분하여 새로운 문서를 분석할 때 그 에러를 최소화하고자 하는 것이 목적이다 (Joachims, 1998).



<그림 1> 지지벡터기계(SVM) 알고리즘

2.2 문서 구성요소와 참조 정보원

문서 범주화와 관련된 연구는 일반적으로 두 가지 범주로 구분될 수 있다. 첫 번째 연구 분야는 전문(full text) 중심의 문서 범주화이다. 전문 중심의 문서 범주화 연구는 주로 효율적인 학

습 알고리즘 개발, 자질 선정 기법, 특정 집합의 문서 범주 성능 최적화 등의 분야에 집중하고 있다 (Cunningham, Witten, & Litten, 1999; Sebastiani, 2002; 2005). 두 번째 분야는 문서 구성요소를 고려한 문서 범주화 연구이다. 이들 연구는 문서 특성과 그 구성요소를 인식하고 그 중요성을 고려하여 문서 범주화 학습 시에 반영하여 문서 범주화 성능 향상을 보이고 있다. 그러나 이들 연구가 특정 문서 구성 요소를 사용했을 때 더 좋은 성능을 나타낸다고 보고하고 있으나, 그 본질적인 과정으로서 문서 구성요소를 색인가가 참조하는 정보원으로 인식하지 못하는 한계가 있다.

이 연구는 두 번째 분야의 문서 범주화 연구와 그 맥을 같이 하고 있으며, 활용가능한 문서 구성 요소를 색인가가 주제 색인 시에 참고하는 정보원으로서 이해하는 데서 그 차이를 두고 있다. 즉, 색인가가 색인 과정 중에 참조하는 문서의 구성 요소를 밝혀 이를 문서 범주화 과정에 반영하여 궁극적으로 문서 범주화의 성능과 효율을 높이고자 하는 것이다. 문서의 구성 요소를 문서 범주화 실험에서 사용하여 성능을 높인 선행 연구들이 있다. 첫째, Larkey(1999)는 kNN 알고리즘을 통해 문서의 제목, 초록, 본문 도입 부분(첫 약 20줄)을 사용하였을 때 가장 문서 범주화 성능이 우수하다고 밝혔다. 이와 유사한 연구로서, Efron, Elsas, Marchionini, & Zhang (2004)은 문서의 키워드와 제목이 본문보다 문서 범주화 성능에서 있어서 효과적이라고 보고하였다. 또한, Zhang, *et al.* (2004)은 인용 서지 정보를 사용하였을 때 문서 범주화 성능에 있어서 효과적이라고 밝혔다. Zhang, *et al.* (2004)의 결과와 유사하게, Slattery (2002)는 지지기반벡터 알고리즘을 사용하여 하이퍼링크 정보가 문서 범주화에서 긍정적인 영향을 미친다고 보고하였다. 보다 심도 깊은 접근으로 Diaz, Ranilla, Montanes, Fernandex, & Combarro (2004)는 문서의 주변 정보(Contextual information) (i.e. 로컬 용어)를 사용하였을 때 보다 효과적인 문서 범주화 성능을 보였다고 보고하고 있다. 이와 같이 초기의 문서 범주화 연구가 일반적으로 전문(full text)을 이용하는 것이 주된 흐름이었으나, 최근 다양해진 문서 집합과 문서 구성 요소에 대한 새로운 인식으로 인해 특정 문서 구성 요소를 이용한 접근이 새로운 연구 흐름으로 부상하고 있다.

3.정보원(Information Resources)

문서 구성 요소를 고려한 문서 범주

화 연구가 새로운 문서 범주화 연구의 흐름으로 인식되고 있으나, 보다 근본적으로 주제 색인 과정과 색인가가 참조하는 정보원으로서의 문서 구성요소에 대한 이해 없이 이루어지고 있는 한계가 있다. 따라서 선행 연구에서 주로 고려되는 문서 구성요소는 제한적이고 근본적인 정보원에 관한 고찰이 없이 실험되고 있는 현실이다. 이 연구는 색인가가 문서의 주제 분석 및 색인어 할당 과정에 대한 이해를 바탕으로 색인가가 색인어를 할당할 때에 참조하는 문서 구성 요소를 문서 범주화를 위한 정보원(Information Resources)으로 파악하고 그 성능을 평가를 하고자 한다.

색인가가 참조하는 정보원에 대한 규명은 일반적으로 세 가지 종류의 문헌에서 찾아 볼 수 있다. 첫째, 주제 색인 시스템과 이에 관한 안내서가 주제 색인을 위한 참고 정보원을 제시하고 있다. Mai (2005)가 밝혔듯이, 듀이십진분류법 (Mitchell, et al., 2003) 에서는 제목, 목차, 소제목, 서문, 서론 등이 주제 분석 및 색인을 위한 정보원이라고 지적하고 있다. 듀이십진분류법과 유사하게 국제표준 (International Standard Organization, ISO 5963:1985) 또한 제목, 초록, 목차, 서문 등이 참고 정보원이라고 밝히고 있다.

둘째, 주제 색인과 주제 목록에 관한 교재에서도 참고 정보원을 밝히고 있다. 우선 Chan (1987)과 Sauperl (2002)은 주제 색인과 주제 목록에 관한 교재를 파악했다. Chan과 Sauperl 파악한 교재들은 Chan (1981), Foskett (1996), Taylor (2003) 등이며, 이들은 공통적으로 제목, 키워드, 인용목록, 서론, 목차 등을 주제 색인 및 주제 목록을 위한 정보원이라고 밝히고 있다. Foskett은 저자, 키워드, 인용목록을 꼽았고, Taylor는 제목/부제목, 목차, 서론 등을 참고 정보원으로 밝혔다. 한편, Chan은 본문과 문서 특질요소 (document attributes)를 비교하여 주제색인 및 목록에 있어서 문서 특질요소를 강조하였다. Chan은 제목, 초록, 목차, 장별 제목, 서문, 서지정보 등 정보원으로 밝혔다.

셋째, 색인가의 실제 색인 작업을 조사 및 분석함으로써 주제 색인 정보원을 파악하고자 하는 일련의 연구가 있다 (Chu & O'Brien, 1993; Jeng, 1996; Sauperl, 2002; 2004). Jeng은 서지정보가 색인가의 주된 주제 색인 정보원이라고 밝히고 있다. Sauperl은 이들 연구 보고는 서지정보, 제목, 저자와 출판사, 부제목, 초록, 소제목, 도입 부문 등이 주제 색인가의 주된 정보원이라고 밝히고 있다. 이와 유사하게 Chu & O'Brien도 제목, 부제목, 초록, 장별 제목, 장별 첫 단락 등이 주제 색인에 있어 중요한 정보원이라고

밝히고 있다.

4. 실험설계

4.1 실험 개요

이 연구는 문서 범주화를 위하여 색인가가 참조하는 정보원을 선별 및 평가하고자 한다. 실험을 위하여서 정보원이 선정되었는데, 선정 기준은 두 가지로 이루어졌다. 첫째는, 선행연구를 통하여 밝혀진 정보원이 우선적으로 고려되었다. 둘째는 밝혀진 정보원 중에서 실험 대상 문서 집합의 특성을 고려하여 다시 선정되었다. 기존 선행연구를 통해 밝혀진 정보원과 저널 및 회의록에 수록된 과학 분야의 논문이라는 데이터 특성이 고려되어 정보원이 선정되었다. 실험을 위하여 최종 선정된 정보원은 제목, 서론, 결론, 참고문헌, 초록, 키워드, 출처이며, 성능 비교를 위하여 본문(full text)도 함께 선정되었다.

4.2 실험 문헌 집단

이 연구는 정보원을 활용하여 문서 범주화 성능을 실험하고자 한다. 따라서 실험 문헌 집단은 다음과 같은 세 가지 조건이 충족되어야 한다. 첫째, 각 문서는 선정된 정보원을 포함하고 있어야 한다. 둘째, 정보원과의 비교를 위하여 전문(full-text) 정보가 구비되어야 한다. 셋째, 색인가에 의하여 주제 색인어가 할당되어야 한다. 이러한 세 가지 전제 조건을 갖춘 문헌집단으로 인스펙트(INSPEC) 데이터베이스가 선정되었다. 인스펙트 데이터베이스는 전자공학, 물리학, 통제과학, 정보과학, 컴퓨터공학, 산업공학 분야의 과학저널의 전문 및 서지정보를 수록하고 있다. 인스팩트 데이터베이스는 8백만 서지정보를 수록하고 있으며 이는 3천 5백 저널과 1천 5백 학회 회의록을 포괄하고 있다 (Engineering Village 2, n.d.). 전형적인 서지정보는 제목, 키워드, 초록, 및 주제 색인어 등이다.

이 연구를 위해 1,000 문서가 문서 집합으로 구성되었으며, 이는 각각 본문과 서지정보를 포함하고 있다. 총 문서는 다시 20개의 주제로 구분되어 각 주제 분야는 50개의 문서를 포함하게 된다. 또한 문서 집합은 두 개의 하위 문서 집합, 즉 동질 문서 집합 (Homogeneous data set)과 이질 문서 집합 (Heterogeneous data set)으로 구성되었다. 동질 문서 집합은 상대적으로 유사한 주제 분야의 문서 집합이며, 이질 문서 집합은 상대적으로 상이한 주제 분야의 문헌으로 구성되었다. 이질 문서 집합은 인스펙트 데이터베이스의 컴퓨터학(Computer Science)

과 정보기술(Information Technology) 전 주제 분야에 걸친 10개의 주제어를 선정하여 검색하였다. 이에 반해 동질 문서 집합은 소프트웨어(Software) 주제 분야에 국한된 10개의 주제어를 선정하여 검색하였다. 동질 문서 집합과 이질 문서 집합으로 구분하여 문서 범주화 실험하는 이유는 다양한 문서 집합 환경 하에서 정보원이 문서 범주화 성능에 영향을 파악하기 위해서이다. 선정된 주제어는 <표 1> 과 같다. 동질 문서 집합은 10개 주제어가 소프트웨어 주제 분야에서 선정되었고, 이질 문서 집합은 컴퓨터학과 정보기술 전반에 걸친 10개의 주제 분야에서 선정되었다.

<표 1> 문서 집합을 위하여 선정된 주제어

동질 문서 집합	이질 문서 집합
software architecture	computer architecture
software development management	computer graphics
software libraries	computer interface
software maintenance	discrete systems
software metrics	information management
software portability	knowledge based systems
software prototyping	pattern recognition
software quality	reliability
software reliability	software engineering
software reusability	user interfaces

문서 집합 구성을 위한 검색과정은 다음과 같이 5단계로 구성되었다. 엔지니어링 빌리지 (Engineering Village 2) 사용자 인터페이스를 통하여 1) 인스팩트 데이터베이스 선택, 2) 상기된 20 주제어로 탐색, 3) 검색 시에 선정된 인스팩트 주제어로 한정, 4) 검색 문서를 영어로 제한, 5) 검색 시에 연도를 2000년부터 2006년 사이로 제한하였다. 연도 제한의 필요성은 인스팩트 데이터베이스의 방대함에서 찾아볼 수 있다. 인스팩트 데이터베이스는 1969년 자료부터 수록하고 있는 방대한 데이터베이스이다. 따라서 오랜 기간의 문서를 포함한 데이터는 학문과 용어 자체의 변화를 반영할 수 있기 때문에 이 연구의 본질적인 목적에 충실하기 위해서 비교적 최근인 2000년부터 2006년도까지의 자료로 제한하여 검색하였다.

4.3 데이터 처리

인스팩트 데이터베이스에 있는 문헌

은 일반적으로 전문(full-text)을 포함하며, 제목, 초록, 키워드, 출처, 인용서지정보 등과 같은 서지정보를 포괄하고 있다. 이 연구의 실험을 위해 데이터 변환과정과 8개의 정보원 추출 과정이 진행되었다. 첫째, 인스펙트 데이터베이스의 전문은 PDF 포맷으로 이루어져 있기 때문에 데이터 변환과정이 필요하다. 즉, 데이터 변환을 위하여 PDF 포맷에서 일반적인 ASCII 텍스트 형식으로 전환하는 과정이 필요하다. 이를 위하여 Adobe Acrobat Capture 소프트웨어가 사용되었다. 둘째, 8가지 정보원을 추출하기 위해 전문(full-text)과 서지정보 모두 사용되었다. 우선, 전문에서 4가지 정보원, 즉 서론, 결론, 참고문헌, 전문 자체를 추출하였다. 서지정보 중에서는 제목, 초록, 키워드, 출처 제목(저널 혹은 학회 회의록 제목)이 추출되었다.

추출된 8가지 정보원은 사전 데이터 처리 과정을 거치었다. 기본적인 처리로는 소대문자 처리, 불용어 처리, 단어 노말라이제이션 과정을 거쳤다. 첫째, 소대문자가 혼합되어 사용된 경우에는 모두 소문자로 변경하였다. 둘째, 문서 집합에서 불용어, 특수기호, 숫자는 제거되었다. 셋째, 단어 변형의 노말라이제이션을 위하여서 Porter (1980)의 스템밍 기법으로 표준화 처리되었다.

문서 범주 성능을 증진시키고 계산적인 복잡성을 줄이기 위해서 문서 범주화 실험의 필수적인 과정인 자질 선정(feature selection)이 고려되었으나 이 연구에서 제한적으로 사용되었다. 이 연구의 목적이 8개의 정보원을 비교하여 문서 범주화 실험에 효율적인 정보원을 밝히고자 하는 것이다. 그러나 정보원의 성격에 따라 <표 2>에서 보는 바와 같이 단어 수에 있어서 현격한 차이를 보인다. 이 중에서 전문(full text)과 함께 전문의 성격이 짙은 서론, 결론, 초록 정보원에 대해서만 자질 선정 과정을 거쳤다. 자질 선정은 기존의 연구에서 성능이 우수한 것으로 밝혀진 정보 획득량 (Information Gain), 문헌빈도 (DF), 카이제곱통계량 (X^2 statistics)을 선정하여 선행 조사하였다. 이 중에서 카이제곱통계량이 가장 우수한 것으로 조사되어 이 연구에서는 카이제곱통계량을 8가지 정보원 중에서 본문 정보원을 비롯하여 본문 성향의 정보원인 서론, 결론, 초록에 적용하였다. 또한 특기할 만한 정보원 데이터 처리는 인용문헌이다. 이 연구의 목적이 인용 분석이 아니기 때문에 인용문헌 중에서 제목만을 추출하여 정보원으로 사용하였다. 마지막으로, 처리된 텍스트는 WEKA (Witten & Frank, 2000) 시스템으로 실험하기 위하여 WEKA 입력 형태인 ARFF (.arff) 파일 포맷으로 변환되었다.

<표 2> 8개 정보원의 단어 수

	전체 데이터	동질 문서 집합	이질 문서 집합
초록	121,699	61,023	60,676
인용문헌	472,985	244,502	228,483
결론	275,253	149,797	125,456
본문	6,602,440	3,507,097	3,095,343
서론	517,197	275,728	241,469
키워드	32,861	15,625	17,237
출처	4,074	2,026	2,048
제목	7,553	3,690	3,863
계	8,034,062	4,259,488	3,774,575

4.4 성능평가

범주화 성능 평가를 위하여 일반적으로 세 가지 측정단위인 재현율, 정확률, F-척도가 주로 사용된다. (Lewis, 1995; Sebastiani, 2002; Yang, 1999) 재현율은 모든 적합 문헌수 중에서 부여된 적합 문헌수로 나타내며, 정확률은 바르게 부여된 문서 수와 틀리게 부여된 문서 주 중에서 부여된 적합 문헌수로 나타낸다. 이러한 재현율과 정확률은 서로 상충되는 부분들이 있어 이를 보완하기 일반적으로 F-척도 (van Rijsbergen, 1979)로서 재현율과 정확률의 평균을 나타낸다.

<표 3> 문서 범주화 성능 평가표

	옳은 주제어	틀린 주제어
옳은 주제어 선정	a	b
틀린 주제어 선정	c	d

재현율 (R) = a / (a + c)

정확률 (P) = a / (a + b)

F 척도 = 2PR / (P + R)

또한 여러 주제 분야 간의 전체적인

성능 평가를 위하여 매크로 평균(Macro-averaging)과 마이크로 평균(Micro-averaging)이 일반적으로 계산 된다 (Yang, 1999). 매크로 평균(Macro-averaging)은 모든 주제 분야에 걸쳐 평균 정확률과 재현율을 계산하며, 마이크로 평균(Micro-averaging)은 각 주제 분야의 평균을 문헌 수에 비례하여 계산하는 방식을 사용한다(Diaz, *et al.*, 2004). 이 연구는 각 주제어 별로 일정한 수의 문헌이 할당되었으므로 매크로 평균과 마이크로 평균이 동일하게 나타난다.

4.5 실험

실험을 위하여서는 자바 프로그램 언어로 구현된 기계학습 어플리케이션인 WEKA (Witten & Frank, 2000) 시스템이 선정되었다. WEKA 시스템은 안정성과 신뢰할 수 있는 성능 구현으로 문서 범주화 연구에 빈번히 사용되어 왔다. WEKA 내에서 지지벡터기계(SVM) 알고리즘 구현은 "weka.classifiers.functions.SMO"으로서 이 연구의 실험을 위하여 사용되었다.

실험은 8개의 정보원으로 구성된 각각의 데이터를 통해 이루어졌다. 실험은 3가지 형태로 구성되었으며 이는 이 중에서 전문(full text)을 사용한 실험을 베이스라인(baseline)으로 파악하여 다른 7개 정보원의 실험결과와 비교 분석하였다. 전문을 사용한 실험결과를 베이스라인으로 인식하는 것은 기존의 문서 범주화 연구가 주로 문서 전문에 치중하였기 때문이다.

5. 실험 결과 및 분석

5.1 문서 집합에 따른 정보원 평가

이 실험은 우선 세 가지 문서 집합(전체 문서 집합, 동질 문서 집합, 이질 문서 집합)에 대하여 문서 범주화 실험을 진행하였다. 실험의 결과는 세 가지 척도인 정확률, 재현율, F-척도로 측정하여졌다. 첫째, 전체 문서 집합을 대상으로 문서 범주화 실험이 이루어졌다. 실험 결과는 <표 4>에서 살펴볼 수 있는 바와 같이 키워드 정보원이 가장 우수한 결과를 나타내었고, 이에 뒤를 이어, 인용, 전문, 출처, 제목 정보원이 비교적 좋은 결과 보이고 있다.

<표 4> 전체 문서 집합 실험 결과

정보원	정확률	재현율	F-척도
초록	0.277	0.234	0.230
인용	0.344	0.290	0.308
결론	0.206	0.193	0.193
전문	0.349	0.283	0.300
서론	0.283	0.213	0.231
키워드	0.387	0.366	0.368

출처	0.323	0.304	0.299
제목	0.319	0.293	0.296

둘째, 동질 문서 집합을 대상으로 한 실험에서는 <표 5>에서 보여주는 바와 같이 전반적으로 전체 문서 집합 실험 결과에 비교하여 미미한 폭으로 상승한 결과를 보여주고 있다. 전체 문서 집합의 결과와 유사하게 키워드 정보원이 가장 우수하고, 인용과 제목 정보원이 그 뒤를 이은 것으로 나타났다.

<표 5> 동질 문서 집합 실험 결과

정보원	정확율	재현률	F-척도
초록	0.283	0.275	0.262
인용	0.345	0.310	0.322
결론	0.249	0.244	0.243
전문	0.287	0.258	0.261
서론	0.286	0.234	0.247
키워드	0.402	0.382	0.384
출처	0.267	0.269	0.262
제목	0.310	0.292	0.295

세 번째로는, 이질 문서 집합을 대상으로 한 실험 결과가 <표 6>에 제시되고 있다. 전체 문서 집합과 동질 문서 집합의 결과와 비교하여 정확률, 재현율, F-척도가 각각 큰 폭으로 상승하였음을 보여 주고 있다. 이 결과는 일반적으로 문서 범주화 성능은 범주의 수가 적을수록 그리고 범주간의 주제상의 거리가 클수록 좋은 결과를 보이는 데 기인한다. <표 6>에서 살펴보는 바와 같이, 키워드 정보원이 가장 우수하며, 키워드 정보원의 뒤를 이어, 전문, 인용, 서론, 출처, 제목 정보원들이 상대적으로 우수한 결과를 보여주었다. 이러한 결과는 전체 문서 집합과 동질 문서 집합의 실험 결과와 유사하다고 판단된다.

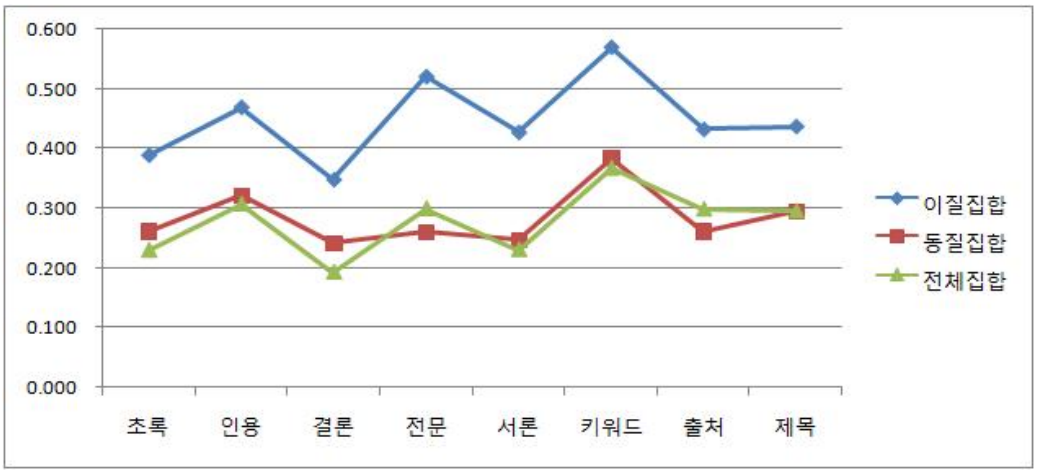
<표 6> 이질 문서 집합 실험결과

정보원	정확률	재현율	F-척도
초록	0.425	0.393	0.389
인용	0.493	0.459	0.469
결론	0.371	0.349	0.348
전문	0.564	0.505	0.520
서론	0.444	0.426	0.427
키워드	0.587	0.568	0.569
출처	0.461	0.420	0.432
제목	0.487	0.422	0.437

이러한 세문서 집합 (전체 문서 집합, 이질 문서 집합, 동질 문서 집합)간의 차이를 좀 더 명확하게 보기 위해 <그림 2>에서와 같이 그래프를 이용하여 시각적으로 비교해 볼 수 있다. 정확율과 재현율의 평균값이라고 볼 수 있는 F-척도를 사용하여 8개의 정보원을 비교하였다. 이들 문서 집합 간의 차이는 이질 문서 집합이 큰 폭으로 우수한 성능을 보이고 있으며, 전체 문서 집합과 동질 문서 집합은 유사한 성능을 나타냈다. 이러한 현상은 일반적으로 문서 범주화 성능은 문서간의 다양성이 확보될수록 성능이 향상되는

선행 연구 결과와 일치한다.

세 가지 문서 집합을 사용한 실험 결과를 종합적으로 살펴 볼 때, <그림 2>에서 보여주는 것과 같이 전문뿐만 아니라 키워드, 인용, 출처, 제목과 같은 정보원이 상대적으로 효율이 높은 정보원으로 파악되었다. 이러한 결과가 어느 한 문서 집합에만 나타나는 제한적인 현상이 아니라 세문서 집합 공통으로 나타나며 실험 결과의 일반화 가능성을 제시하고 있다.



<그림 2 > 전체, 이질 동질 문서 집합 성능 비교 (F-척도)

5.2 정보원간 비교 평가

정보원간의 성능을 비교하기 위하여 전체 문서 집합, 동질 문서 집합, 이질 문서 집합의 F-척도 값에 대하여 *t*-검증을 사용하여 비교하였다. 이를 통하여, 8개의 정보원의 문서 범주화 실험에서의 각각의 정보원과 비교하여 상대적인 효율성을 파악하고자 하였다.

첫째, <표 7>에서 살펴볼 수 있는 바와 같이 전체 문서 집합을 대상으로 전문(full text)과 7개 정보원을 *t*-검증을 사용하여 비교하였다. 신뢰도 구간을 95%로 하였을 때 다수의 정보원이 다른 정보원과 비교하였을 때 유의한 차이를 보이고 있다. 우선, 인용 정보원은 결론과 서론 정보원에 비교하여 유의한 차이가 있으며, 키워드 정보원은 초록, 결론, 서론 정보원들과 서로 유의한 차이를 보이고 있다. 저널 제목 정보원은 초록, 결론, 키워드 정보원들과 유의한 차이를 보이며, 제목 정보원은 초록, 결론, 서론, 키워드 정보원들과 유의한 차이를 보이고 있다.

둘째, <표 7>에서 제시된 바와 같이 동질 문서 집합을 이용하여 정보원 간의 성능을 *t*-검증을 이용하여 비교 분석하였다. 전체 문서 집합과 비교하였을 때 정보원 간의 차이가 큰 폭으로는 두드러지지 않은 것으로 나타났다. 가장 특기할 만한 특징은 키워드 정보원에 대한 부각이다. 즉, 키워드 정보원은 결론, 서론, 초록, 출처, 제목 정보원과 비교하여 유의한 차이를 보이고 있다. 이는 전체 문서 집합과 비교하여 동일하게 나타난 현상이다.

<표 7> 정보원간의 F-척도의 *t*-검증 결과 (동질 문서 집합)

	초록	인용	결론	서론	키워드	출처	제목
초록		0.357	0.520	0.782	**0.000**	0.993	0.431
인용			0.243	0.416	0.283	0.369	0.692
결론				0.940	**0.001**	0.648	0.221
서론					**0.044**	0.811	0.357
키워드						**0.001**	**0.010**
출처							0.381
제목							

$p<0.05$

셋째로는, 이질문서집합은 <표 8>에서 제시하는 바와 같이 8가지 정보원들이 서로 유의한 차이가 있음을 보이고 있다. 우선, 키워드 정보원은 초록, 결론, 서론, 출처, 제목 정보원과 유의한 차이를 보이고 있다. 이는 전체 문서 집합과 동질 문서 집합과 동일하게 나타난 현상이다. 또한 특기할 만한 사항은 인용 정보원은 결론 정보원과 유의하게 차이를 나타내고 있으며, 제목 정보원과 결론 정보원이 서로 유의하게 차이를 나타내는 것이다. 각 문서 집합에서 정보원간에 유의한 차이를 보이는 것은 문서 범주화에 미치는 영향에 있어서 정보원이 동일하지 않다는 것을 나타낸다고 볼 수 있다.

<표 8> 정보원간의 F-척도의 *t*-검증 결과 (이질 문서 집합)

	초록	인용	결론	서론	키워드	출처	제목
초록		0.148	0.335	0.270	**0.006**	0.441	0.346

인용			**0.001**	0.170	0.069	0.526	0.544
결론				**0.002**	**0.001**	0.120	**0.047**
서론					**0.001**	0.914	0.732
키워드						**0.004**	**0.001**
출처							0.879
제목							

*$p<0.05$

5.3 정보원과 전문 간의 비교 평가

이 연구의 목적이 문서 범주화 효율성 제고를 위하여 8가지 정보원의 성능을 평가하는데 있으며, 특히 기존의 연구에서 참작하지 못한 색인가 참조 정보원 즉, 제목, 인용, 초록, 키워드 등의 문서구성요소의 효율성을 살펴보고 그 중요성에 대한 인식을 제고 하는데 있다. 그러므로 전문(full-text)을 베이스라인으로 하여 다른 7가지 정보원을 비교 평가하는 것이 의미가 있다. <표 9>에서 제시하는 바와 같이 신뢰도 구간을 95%로 하였을 때 F-척도상의 유의한 차이를 보이는 정보원이 발견되었다. 전체 데이터 셋에서는 초록, 결론, 서론, 키워드 정보원이 본문 정보원과 비교하여 유의한 차이를 보였다. 동질 문서 집합에서는 키워드 정보원이 본문 정보원과 유의한 차이를 보였다. 한편으로 이질 문서 집합에서는 초록과 결론 정보원이 본문 정보원과 비교하여 유의한 차이를 보이고 있다.

<표 9> 전문과 정보원간 F-척도 t-검증

정보원	전체문서집합		동질문서집합		이질문서집합	
	t	p	t	p	t	p
초록	3.446	**0.003**	-0.025	0.981	3.273	**0.010**
인용	-0.249	0.806	-0.857	0.414	1.252	0.242
결론	4.481	**0.000**	0.500	0.629	4.333	**0.002**
서론	2.587	**0.018**	0.310	0.763	2.275	0.051
키워드	-2.231	**0.038**	-5.212	**0.001**	-0.753	0.471
저널 제목	0.043	0.966	-0.015	0.988	1.402	0.194
제목	0.110	0.913	-1.267	0.237	1.321	0.219

*$p<0.05$

한편, <표 10>에서 제시된 바와 같이 재현율에 있어서는 F-척도와 비교하여 적은 수의 정보원이 베이스라인과 비교하여 유의한 차이를 보이는 것으로 나타났다. 전체 데이터 셋에서는 결론과 키워드 정보원이 전문 정보원과 비교하여 유의한 차이를 보였다. 동질 문서 집합에서는 유일하게 키워드 정보원만이 유의한 차이를 보였으며, 이는 F-척도와 동일하다. 또한, 이질 문서 집합에서는 결론 유일하게 결론 정보원만이 유일하게 유의하게 차이를 보이고 있다.

<표 10> 전문과 정보원간 재현율 *t*-검증

정보원	전체문서집합		동질문서집합		이질문서집합	
	t	*p*	*t*	*p*	*t*	*p*
초록	1.443	0.165	-0.441	0.669	1.763	0.112
인용	-0.160	0.875	-0.601	0.563	1.398	0.196
결론	3.432	**0.003**	0.300	0.771	4.176	**0.002**
서론	1.787	0.090	0.339	0.742	1.619	0.140
키워드	-2.935	**0.008**	-3.945	**0.003**	-1.057	0.318
저널 제목	-0.852	0.405	-0.852	0.405	1.667	0.128
제목	-0.354	0.727	-0.354	0.727	1.733	0.117

*$p<0.05$

마지막으로 베이스라인인 전문과 정확률 간의 *t*-검증 결과가 <표 11>에서 제시되었다. 먼저, 전체 문서 집합을 살펴보면, 결론과 서론 정보원이 전문 정보원과 유의한 차이를 보이고 있다. 동질 문서 집합에서는 유일하게 키워드 정보원만이 전문 정보원간의 유의한 차이를 보이고 있다. 이질 문서 집합에서는 초록과 결론 정보원이 모두 전문과 유의한 차이를 보이고 있다.

<표 11> 전문과 정보원간 정확률 *t*-검증

정보원	전체문서집합		동질문서집합		이질문서집합	
	t	*p*	*t*	*p*	*t*	*p*
초록	2.356	0.029	0.120	0.907	2.306	**0.047**
인용	0.131	0.897	-0.925	0.379	1.096	0.302
결론	4.523	**0.000**	1.086	0.306	2.611	**0.028**
서론	2.389	**0.027**	0.010	0.992	2.034	0.073
키워드	-0.924	0.367	-4.757	**0.001**	-0.286	0.781

저널 제목	0.704	0.490	0.476	0.645	1.235	0.248
제목	0.644	0.527	-.510	0.623	0.938	0.373

*p<0.05

6. 결론

대량의 문서에 주제어를 자동으로 할당하는 문서 범주화(Text Categorization) 연구가 최근 활용 가능한 전자 문서와 컴퓨터 하드웨어 및 소프트웨어 기능 향상으로 인해 활발하게 진행되고 있다. 이러한 문서 범주화 연구는 주로 전문(full text) 중심으로 이루어져 왔다. 이는 전산학 분야에서 주도한 연구 접근 방식으로 인해 전문을 이용한 자질 선정, 알고리즘 개발, 최적화 등의 분야에 집중되었다. 이 연구는 전문 중심의 문서 범주화 연구와 달리 색인가가 색인어를 할당하는 과정에 대한 인식에서 출발한다. 특히, 색인과정에서 색인가가 참조하는 문서구성요소를 파악한 후에 이들을 문서 범주화의 성능 효율에 영향을 줄 수 있는 정보원으로 보고자 한 것이다. 파악된 정보원들과 전문(full text)을 이용한 문서 범주화의 성능 측면에서 비교하였다. 이 연구는 과학 기술 분야의 저널 및 회의록 기사를 문서 집합으로 하여, 문서 집합 특성에 맞는 정보원을 선정하여 문서 범주화 성능에 미치는 영향을 실험하여 보고하였다. 이들 정보원은 기존의 연구에 집중하였던 본문(full text) 및 본문 성향의 정보원인 서론과 결론과 함께, 문서 구성 요소인 제목, 출처, 인용, 초록, 키워드를 포함하였다. 한편 문서 집합은 전체 문서집합, 동질 문서 집합, 이질 문서 집합과 같이 다양한 문서 집합으로 구성하였다. 이러한 다양한 문서 집합의 적용은 실험결과의 타당성을 높이기 위함이다.

성능 분석을 위해 세 가지로 구분하여 실험하였다. 첫째, 일반적 성능 분석을 위하여 F-척도로 살펴보면, 인용, 전문, 키워드가 상위그룹의 성능을 보이고 있다. 이러한 결과는 세문서 집합 모두에서 찾아볼 수 있는 공통적인 현상이다. 특히, 인용과 키워드가 좋은 성능을 보이는 것은 Efron, Elsas, Marchionini, & Zhang (2004), Zhang, et al. (2004), Slattery (2002) 등의 연구 결과와도 일치한다. 이러한 결과는 전문 이외에도 문서 범주화에 사용될 수 있는 정보원을 발굴할 수 있는 가능성을 보여주는 것이라고 볼 수 있다. 둘째, 정보원간의 차이를 살펴보기 위해 t-검증을 이용하여 정보원간의 비교분석을 실행했다. 이 분석을 통하여서 키워드 정보원이 다

른 정보원들과 비교하여 유의한 차이를 나타냈다. 이는 문서 집합의 특성에 따라 선정된 정보원은 성능에 있어서 그 차이가 있음을 나타내며, 정보원 선정 시에 활용될 수 있다. 마지막으로, 각 정보원과 본문의 성능을 t-검증을 사용하여 비교하였다. 이 중에서 인용, 출처, 제목 정보원은 본문 정보원과 비교하였을 때 유의한 차이를 보이지 않았다. 이러한 결과는 문서 범주화에 있어서 본문 보다는 전처리 과정 및 처리 속도 면에서 효율적인 정보원들이 발굴될 수 있음을 나타낸다. 이러한 정보원을 선택 활용하는 것이 문서 범주화에 있어 효율적이라는 가능성을 제시하고 있다. 본문 중심의 문서 범주화는 자질 선정 및 가중치 적용 등의 사전 가공 처리 과정이 필요하며, 최적화된 상태를 먼저 파악해야 한다는 부담이 있기 때문이다. 또한 키워드 정보원은 다른 정보원과의 비교에서 나타난 바와 같이 본문 정보원과 비교하여 유의한 차이를 보이고 있다. 이러한 결과는 키워드 정보원은 문서 범주화에 있어서 가장 효과적인 결과를 낼 수 있는 중요한 정보원으로 인식된다. 따라서 문서의 성격에 따라 키워드 정보원이 활용 가능하다면, 본문 정보원 보다는 키워드 정보원 활용이 바람직한 것으로 여겨진다.

참고문헌

Chan, L.M. (1981). Cataloging and classification: An introduction. New York City, NY: McGraw-Hill.

Chan, L.M. (1987). Instructional materials used in teaching cataloging and classification. *Cataloging and Classification. 7*, 131-144.

Chu, C.M. & O'Brien, A. (1993). Subject analysis: The critical first stage in indexing. *Journal of Information Science. 19*, 439-454.

Cunningham, S.J., Witten, I.H., & Littin, J. (1999). Applications of machine learning in information retrieval. *Annual Review of Information Science and Technology*, 34, 341-384.

Diaz, I., Ranilla, J., Montanes, E., Fernandez, J., & Combarro, E. (2004). Improving performance of text categorization by combining filtering and support vector machines, *Journal of the American Society for Information Science and Technology, 55*(7), 579-592.

Efron, M., Marchionini, G., Elsas, J., & Zhang, J. (2004). Machine learning for information architecture in a large governmental website. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries,* 151-159.

Engineering Village 2. (n.d.). Retrieved November 11, 2006, from http://www.engineeringvillage2.org/controller/servlet/Controller.

Foskett, A.C. (1996). *The Subject Approach to Information.* London: Library Association Publishing.

ISO 5963: 1985. (1985). *Documentation-methods for examining documents: Determining their subjects and selecting indexing terms.* International Standards Organization.

Jeng, L.H. (1996). Using verbal reports to understand cataloging expertise: Two cases, *Library Resources and Technical Services,* 40(4), 343-358.

Joachims, T. (1998). Text categorization with support vector machine: Learning with many relevant features,

Proceedings of the 10th European Conference on Machine Learning, 137-142.

Larkey, L.S. (1999). A patent search and classification system. *Proceedings of the 4th ACM Conference on Digital Libraries,* 179-187.

Lewis, D.D. (1995). Evaluating and optimizing autonomous text categorization systems. Unpublished Doctoral Dissertation, University of Massachusetts, Massachusetts.

Mai, J.E. (2005). Analysis in indexing: document and domain centered approaches, *Information Processing and Management,* 41, 599-611.

Mitchell, J.S. et al. (Eds.). (2003). *Dewey Decimal Classification and Relative Index*. Dublin, OH: OCLC Online Library Computer, Inc.

Moens, M.F. (2000). *Automatic Indexing and Abstracting of Document Texts.* Norwell, MS: Kluwer Academic Publishers.

O'Connor, B.C. (1996). *Explorations in Indexing and Abstracting: pointing, virtue, and power*. CO: Libraries Unlimited.

Porter, M.F. (1980). An algorithm for suffix stripping, *Program, 14*, 130-137.

Sauperl, A. (2002). *Subject determination during the cataloging process*. Lanham, MD; Scarecrow Press.

Sauperl, A. (2004). Catalogers' common ground and shared knowledge. *Journal of the American Society for Information Science and Technology, 55*(1), 55-63.

Sebastiani, F. (2002). Hypertext categorization. In A. Zanasi (Eds.), *Text Mining and Its Applications* (pp. 109-129), Southhampton, U.K.: WIT Press.

Sebastiani, F. (2005). Text categorization. In A. Zanasi (Eds.), Text mining and its applications (pp. 109-129), Southhampton, U.K.; WIT Press.

Slattery, S. (2002). *Hypertext categorization*. Unpublished Doctoral Dissertation. School of Computer Science. Carnegie Mellon University, Pittsburgh, PA.

Taylor, A.G. (2003). *The*

organization of information (2nd ed.). Englewood, CO; Libraries Unlimited.

van Rijsbergen, C.J. (1979). *Information Retrieval.* Butterworths, London.

Witten, I.H. & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations.* CA: San Diego, Academic Press.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1, 69-90.

Zhang, B., Goncalves, M.A., Fan, W., Chen, Y., Fox, E.A., Calado, P. & Cristo, M. (2004). Combining structural and citation-based evidence for text categorization, *Proceedings of the 13th ACM Conference on Information and Knowledge Management, 162-163.*