자동색인을 위한 학습기반 주요 단어(핵심어) 추출에 관한 연구*

Learning-based Automatic Keyphrase Indexing from Korean Scientific LIS Articles

김혜진, 연세대학교 문헌정보학과, erin.hj.kim@yonsei.ac.kr
정유경, 연세대학교 문헌정보학과, yk.jeong@yonsei.ac.kr
Hea-Jin Kim, Department of Library and Information Science, Yonsei University
Yoo-Kyung Jeoung, Department of Library and Information Science, Yonsei University

학술 데이터베이스를 통해 방대한 양의 텍스트 데이터에 대한 접근이 가능해지면서, 많은 데이터로부터 중요한 정보를 자동으로 추출하는 것에 대한 필요성 또한 증가하였다. 특히, 텍스트 데이터로부터 중요한 단어나 단어구를 선별하여 자동으로 추출하는 기법은 자료의 효과적인 관리와 정보검색 등 다양한 응용분야에 적용될 수 있는 핵심적인 기술임에도, 한글 텍스트를 대상으로 한 연구는 많이 이루어지지 않고 있다. 기존의 한글 텍스트를 대상으로 한 핵심어 또는 핵심어구 추출 연구들은 단어의 빈도나 동시출현 빈도, 이를 변형한 단어 가중치 등에 근거하여 핵심어(구)를 식별하는 수준에 그쳐 있다. 이에 본 연구는 한글 학술논문의 초록으로부터 추출한 다양한 자질 요소들을 학습하여 핵심어(구)를 추출하는 모델을 제안하였고 그 성능을 평가하였다.

1. 서 론

방대한 문서집합으로부터 자동으로 핵심어를 추출하는 기법은 텍스트 마이닝의 한 분야로, 문헌의 주제를 대표하는 대표어(representative term) 또는 주요어(keyphrase or keyword)를 추출하는 것을 그 목적으로 한다. 추출된 핵심어는 이용자에게 문헌에 대한 이해를 높이고, 수작업 색인에 비해 시간과 비용 면에서 효율적일 뿐만 아니라 정보검색, 문헌분류, 요약, 주제탐지 등의 다양한 분야에서 활용될 수 있는 핵심 기술이다.

과거에는 학술문헌의 초록이나 전문과 저자가 부여한 키워드를 함께 수집하는데 어려움이 있었으나, 인터넷 미디어의 발달과 온라인 문서에 대한 접근이 수월해지면서, 다양한 메타데이터를 포함한 학술문헌을 학습문헌집단으로 활용할 수 있게 되었다. 그럼에도 불구하고 핵심어 추출 관련 국내연구는 많은 연구들이 저자 키워드의 활용을 배제한 체, 형태소 분석을 활용한 언어학적 기법이나 문헌 내 단어의 출현정보를 활용한 통계적 기법에 의존을 하고 있다(김지숙 외, 2001; 신성윤, 이양원, 2009; 이성직, 김한준, 2009; 한승희, 2010).

본 연구에서는 Witten et al.(1999)이 제시한 핵심어(구) 추출 알고리즘(Keyphrase Extraction Algorithm, 이하 KEA)을 한글문서에 적용하여 한글문헌대상 핵심어 추출 모델을 생성하는데 목적이 있다. 저자 키워드가 부여된 한글문헌을 대상으로 학습문헌집단을 구축하고 핵심어 추

* 이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015S1A3A2046711).

출 모델을 생성하였다. 본 연구에서는 저자가 문헌의 핵심어로 인식하고 직접 부여한 키워드가 문헌을 가장 잘 대표하는 주요어라는 가정을 바탕으로, 이를 활용한 지도학습기반의 핵심어 추출 모델링을 수행하였다. 이는 저자의 의도가 담긴 주요어를 문헌에서 직접 추출하는 효과적인 모델이 될 수 있다고 판단하였기 때문이다.

2. 선행연구

기존의 핵심어 추출은 주로 단어의 빈도 정보를 활용하거나 빈도 가중치를 변형한 통계기반 기법과 품사 및 구문분석 등을 통해 언어학적 자질을 활용하는 언어학적 접근기법, 통계적 기법과 언어학적 기법을 기반으로 문헌내의 단어 동시출현 정보를 활용하여 후보 핵심어를 추출하는 비지도학습기반 접근기법으로 나누어 볼 수 있다. 이성직과 김한준(2009)은 기존의 TFIDF 모델을 활용하여 분야별 뉴스기사의 후보 키워드를 추출하였다. 뉴스기사에서 일반적으로 등장할 수 있는 매체이름과 기자이름 등 출현빈도는 높으나 불용어로 처리되어야 할 단어들을 걸러내기 위해서 분야 간 교차비교 분석을 수행하여 의미 없는 단어를 제거하였다.

저자 키워드가 부여된 학습문헌집단의 구축이 어려운 경우, 문헌내의 단어들의 동시출현 정보를 바탕으로 단어들 간의 관계를 이용한 한 비지도학습 기반 핵심어 추출기법이 활용하였다. 신성윤과 이양원(2009)은 형태소 분석을 사용하여 추출한 명사를 대상으로 연관규칙 탐사 알고리즘을 적용한 비지도학습 기반 키워드 추출기법을 제안하였다. 김지숙 외(2001)도 컴퓨터분야의 논문을 대상으로 세부주제를 대표하는 특정단어집합을 씨앗단어(seed keyword)로 선정한 후 연관규칙탐사 알고리즘을 적용하여 대표 색인어 추출하였다. 이들이 사용한 연관규칙 탐사 알고리즘을 적용한 특정 분야의 핵심색인어 추출은 통계적 기법을 적용한 X^2(Chi square)기법과 DF (Document Frequency)기법보다 문헌을 더 효율적으로 표현할 수 있었다. 한승희(2010)는 용어의 출현빈도를 기반으로 용어클러스터링을 이용하여 문헌의 핵심어를 자동 추출하였다. 추출한 키워드를 좋은 키워드의 전제조건이라고 할 수 있는 주제성과 고른 빈도분포(중빈도어) 측면에서 평가했을 때 우수한 결과를 보여주었다.

그러나 저자가 문헌의 핵심어로 인식하고 직접 부여한 키워드가 문헌을 가장 잘 대표하는 주요어임에도 불구하고 이를 활용하여 핵심어를 추출한 연구가 국내에서는 거의 전무한 실정이다. 저가 키워드를 활용한 지도학습기반 핵심어 추출 문헌 모델링은 저자의 의도가 담긴 주요어를 문헌에서 직접 추출하는 효과적인 모델이 될 수 있을 것이다.

3. 학습기반 핵심어 추출

3.1 데이터 수집

핵심어 추출 모델의 학습데이터 구축을 위해, 국내 문헌정보학분야의 저널 4종(정보관리학회지, 한국도서관·정보학회지, 한국문헌정보학회지, 한국비블리아학회지)에 대하여 국내저널 데이터베이스인 DBPIA(www.dbpia.co.kr)에서 제공하는 모든 기간의 논문을 수집하였다. 이중 키워드가 제공되지 않은 논문을 제외한 총 2,905건의 문헌을 대상으로 학습문헌집단을 구축하였다(표 1).

<표 1> 수집데이터 및 학습문헌집단 통계

학회지명	수집기간	문헌수
정보관리학회지	2006.12-2017.6	700
한국도서관·정보학회지	2006.12-2016.9	830
한국문헌정보학회지	2006.12-2017.5	805
한국비블리아학회지	2006.12-2017.6	570
총 문헌수		2,905

3.2 핵심어 추출 알고리즘

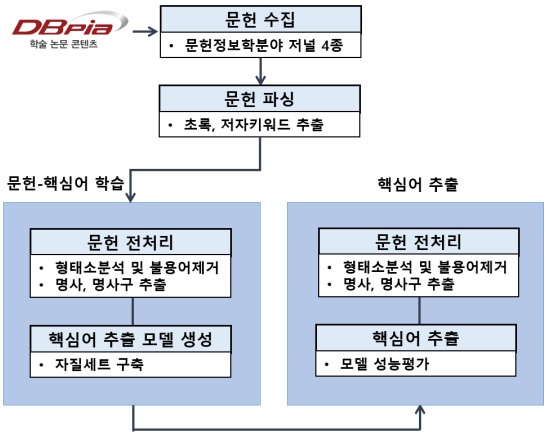
KEA는 저자키워드를 가지고 있는 학습문헌집단을 학습하여 핵심어 추출 모델을 생성하는 단계와 새로운 문헌이 입력되었을 때 생성한 모델을 기반으로 핵심어를 추출하는 단계로 구성된다. 문헌 내의 핵심어(구)를 추출하기 위해서 다음의 네 가지 자질을 활용한다(Witten et al., 1999).

① TF·IDF: 전체 문헌집단 내에서의 단어의 특정성(specificity)
② 출현정보: 핵심어가 문헌 내에서 처음으로 등장하는 출현정보
③ 핵심어(구) 길이: 핵심어 또는 핵심어 구를 구성하는 단어의 수
④ 연결정도: 핵심어와 의미적으로 연결된 단어의 수

3.3 핵심어 추출 모델 구축

KEA는 문헌이 입력되었을 때, 문자가 아닌 단어들과 일반적인 형용사와 부사, 동사를 포함한 불용어 리스트와 대조하여 이런 단어들은 핵심어 후보군에서 제외한다. 영어의 특성상 항상 대문자로 시작하는 고유명사도 추적하여 후보군에서 제외하였다. 따라서 본 연구에서도 학습문헌과 검증문헌에서 한글 형태소분석기를 사용하여 문헌을 전처리하였고, 구두점, 조사, 대명사 등은 불용어 처리하여 핵심어의 후보군에서 제외하였다. 한글 전처리를 위해서 사용한 형태소분석기는 코모란 2.4버전(www.shineware.co.kr)을 활용하였다.

본 연구의 핵심어 추출 모델생성 및 성능평가는 <그림 1>과 같다. DBPIA를 통해 수집된 문헌정보학관련 학술논문 총 2,905건을 대상으로, 초록과 저자 키워드를 추출하는 문헌 파싱의 단계를 거쳤다. 이후 한글 형태소분석기를 사용하여 문헌을 전처리하고, KEA의 핵심어 추출 알고리즘을 사용하여 문헌의 자질세트 구축하고 핵심어 추출 모델을 생성하였다. 실험 문헌집단을 대상으로 생성된 모델을 사용하여 핵심어 추출 모델의 성능을 평가하는 과정을 거쳤다.



<그림 1> 핵심어 추출 모델 생성 및 검증

4. 핵심어(구) 추출 실험 결과

핵심어(구) 추출을 위한 학습모델의 성능평가를 위하여, 추출된 키워드와 실험 대상문헌에 부여된 저자 키워드를 비교하여 공통으로 추출된 핵심어의 수를 측정하였다. 실험 대상문헌에 부여된 저자 키워드는 일반적으로 두 단어 혹은 세 단어로 구성되어 있는 경우가 많기 때문에, 성능평가를 위하여 최대로 추출할 수 있는 핵심어구의 단어 수를 두 단어, 세 단어로 구성하여 실험을 진행하였다. 결과는 다음 <표 2, 3>과 같다.

<표 2> 핵심어(구) 추출 성능 (두 단어이하)

추출된 키워드 수	저자 키워드와 매치되는 키워드 수 평균
1	0.336
5	0.911
10	1.120
15	1.199
20	1.228

추출할 수 있는 핵심어를 두 개의 단어로 제한하여 추출하였을 경우, 저자 키워드와 매치되는 키워드의 평균은 20개였을 때 가장 좋은 성능을 보였다.

<표 3> 핵심어(구) 추출 성능 (세 단어이하)

추출된 키워드 수	저자 키워드와 매치되는 키워드 수 평균
1	0.341
5	0.973
10	1.221
15	1.335
20	1.385

핵심어를 추출 시, 핵심어를 구성하는 단어의 수를 최대 세 개까지 지정했을 경우에도 추출하는 핵심어와 저자 키워드와 매치되는 평균은 20개였을 때 가장 좋은 성능을 나타냈다.

핵심어 추출 모델의 성능이 평균 1개 내외로 나타난 이유는 실험문헌집단에서 한 문헌에 부여된 저자 키워드 약 4.5개 중, 실험문헌에 부여된 저자 키워드의 상당수가 추출대상 문헌에 출현하지 않은 단어들이기 때문이다. 또한 저자 키워드와 일지하지 않았던 핵심어(구)들은 해당 문헌에 많이 출현한 일반적인 단어로, 저자 키워드와는 일치하지는 않았으나 해당 문헌과 관련된 핵심어들이었다.

5. 결 론

본 연구는 KEA를 한글 문서에 적용하여 핵심어를 추출할 수 있도록 저자 키워드가 부여된 한글문헌을 대상으로 학습문헌집단을 구축하고 한글기반의 핵심어 추출 모델을 생성하였다. 본 연구에서 생성한 한글 핵심어 추출 모델을 활용하면 복잡한 단어가중치의 계산이나 문헌의 전처리 없이 대용량 전자문서로부터 핵심어(구)의 추출이 가능하다. 추출된 핵심어들은 텍스트 마이닝을 활용한 문서 브라우징, 주제탐지, 자동분류, 정보검색 시스템 등에 적용될 수 있을 것이다.

보다 정확한 한글 핵심어 추출을 위해서는 한글 전처리의 성능 개선이 필요하며, 핵심어 추출의 오류 개선을 위하여 정확률, 재현율 등 다양한 평가지표를 사용하여 추출된 단어들에 대한 오류분석이 수행되어야 할 것이다.

참고문헌

김지숙, 김영지, 문현정, 우용태. (2001). “효율적인 문서 자동 분류를 위한 대표 색인어 추출 기법”. 정보기술과 데이터베이스 저널, 8(1), 117-128.

신성윤, 이양원. (2009). “한국어 정보처리를 위한 명사 및 키워드 추출”. 한국컴퓨터정보학회논문지, 14(3), 51-56.

이성직, 김한준. (2009). “TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법”. 한국전자거래학회지, 14(4), 59-73.

한승희. (2010). “용어 클러스터링을 이용한 단일문서 키워드 추출에 관한 연구”. 한국문헌정보학회지, 44(3), 155-173.

Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., & Nevill-Manning, C.G. (1999). “KEA: Practical automatic key phrase extraction”. In Proceedings of the fourth ACM conference on Digital libraries (pp. 254-255).