효율적인 검색을 위한 논문 키워드 추출 알고리즘 설계 및 연구 검색 시스템 개발

이종현, 이원준, 김호숙
한국과학영재학교
E-mail: jonghyun777@gmail.com, wjl0316@naver.com, khosook@kaist.ac.kr

Academic Paper Keyword Extracting Algorithm for Efficient Search and Development of Research Searching System

Jong-Hyun Lee, Won-Joon Lee, Ho-Sook Kim
Korea Science Academy of KAIST

요 약

본 연구는 논문을 기반으로 연구의 주요 키워드를 추출하는 알고리즘을 설계하고, 이를 적용한 연구 검색 시스템을 개발하여 효율적인 검색 환경을 제공하는 것을 목표로 한다. 논문 키워드 추출 알고리즘은 논문 내에서의 단어 출현 빈도와 PMI 지표를 바탕으로 정의한 단어간 연관성 K(x,y)을 기반으로 설계하였다. 연구 검색 시스템은 고등학교 R&E 등 제한적인 환경에서 이루어지는 연구들의 선행 연구 자료 부족을 해결하는 것을 주 목적으로 한다. 또한, 구현한 연구 검색 시스템에 제안된 알고리즘을 적용하여 보다 정확하고 직관적인 검색 환경을 제공할 수 있었으며, 추후 연구 자료가 추가됨에 따라 그 가치가 높아질 것으로 전망한다.

1. 서론

선행 연구 조사는 연구를 진행하기 이전 과정에서 필수적인 요소이다. 기존 연구 사례들을 참고함으로써, 연구 환경에 부합하는 적합한 범위의 주제와 결과물, 해당 분야와 관련된 연구를 진행한 연구자와 지도교수의 정보 등을 얻을 수 있다. 특히 고등학교에서 이루어지는 R&E 의 경우 연구 환경이 상당히 제한적으로, 연구 내용 자체뿐만 아니라 지도교사 선정 등 부가적인 요소도 필요하기 하기 때문에 선행 연구 조사의 필요성이 매우 높다.

그러나 한국과학영재학교의 경우, R&E와 졸업 연구를 합하여 연 150여건의 연구가 이루어지는 반면, 교내에 보관되는 책자형 논문집 이외에는 연구 기록물을 제공하고 있지 않다. 이에 연구 검색 시스템을 구축하여 신규 연구자의 경우 검색과 접근이 용이한 선행 연구 자료로써, 기존 연구자의 경우 연구 결과 기록물로써 활용할 수 있도록 하고자 하였다.

검색과 접근의 용이성을 위해서는 합리적이고 직관적인 기준이 필요하다. 연구의 제목이나 연구자가 작성한 주제어만으로는 검색어와의 연관성이 떨어지는 명확한 한계가 존재하기에 연구 내용을 표현할 수 있는 키워드의 필요성을 체감하였다. 이러한 키워드의 경우 모든 연구에 대하여 수동으로 생성하는 것은 사실상 불가능하며, 배경지식이 없는 경우 부정확할 가능성이 있다. 이에 본 연구는 논문에서 추출하는 자동화 알고리즘을 개발하고, 이를 연구 검색 시스템에 적용하는 것을 목적으로 한다.

2. 선행 연구 및 배경 지식

2.1. 한국어 형태소 분석

한국어 텍스트의 키워드 추출을 위해서는 단어의 여러 형태들을 하나로 통일시킬 필요가 있다. '많음', '많았다', '많은' 등의 단어들은 모두 '많다'가 기본형이듯, 문장 속 특수한 상황에 따라서 여러 형태로 나타난다. 정확한 텍스트 분석 및 키워드 추출을 위해서는 이러한 단어들을 하나의 뜻으로 인식해야 하기 때문에, 각 단어들의 형태소를 분석하는 작업은 필수적이다.

형태소란 '의미를 가지는 가장 작은 말의 단위'를 의미하며 어휘적 및 문법적 의미를 모두 포함하는 개념이다.[1] 예를 들어서 '우리나라 학생들은 공부를 열심히 한다'의 형태소를 분석해보면 '우리(대명사) + 나라(명사) + 학생(명사) + 들(접미사) + 은(보조사) + 공부(명사) + 를(조사) + 열심히(부사) + 한다(동사)'가 된다. 이처럼 다양한 종류의 형태소 중, 단어가 표현하는 의미와 상관이 없는 형태소들을 미리 제거하게 된다면, 정확도를 높이고 소요

[1] https://ko.wikipedia.org/wiki/%ED%98%95%ED%83%9C%EC%86%8C

시간을 줄일 수 있다.

형태소 분석을 통해서 키워드 추출을 실행한 선행 논문들을 보면 체언만 이용한 경우와[1], 명사와 형용사를 함께 이용한 경우가 있다[2]. 본 연구에서는 연구의 키워드를 추출함에 있어 '암호화','행렬화' 등 동사로 표현되는 단어들도 중요하다고 생각하여 체언, 동사, 형용사를 바탕으로 키워드 추출을 구현하였다. 이 연구에서는 Python 한국어 형태소 분석 패키지인 KoNLPy[2]의 모듈 중 Mecab Class 를 이용해서 형태소를 분석하였다.

2.2. PMI 를 기반으로 한 단어간 연관성 K(x,y)

PMI(Probability Mutual Information)는 통계학과 정보이론에서 쓰이는 지표로, 확률론에 기초하여 두 확률 변수의 특정한 값의 연관성을 표현한다. 서로 연관 되어있는 의미를 가진 단어들은 같이 나타날 확률이 높다는 사실에 입각할 때, PMI 지표로 어휘의 연관성을 측정할 수 있으며[3], 두 어휘의 연관성은 아래와 같은 식으로 도출할 수 있다.

$$\mathrm{PMI}(x,y) = \log\frac{p(x,y)}{p(x)p(y)}$$

$p(x)$: x의 출현 빈도; $p(x,y)$: x,y의 동시 출현 빈도

두 단어가 자주 인접하게 나타나는 등 연관성이 높다면 PMI 값은 커질 것이고, 그렇지 않다면 작아질 것이다.

이 방법의 경우 단순히 단어의 출현 빈도 자체에 의존하는 것이 아닌, 단어쌍과 단어의 출현 빈도 비율에 의존하기 때문에, 전체 출현 횟수가 1 회인 단어가 한 쌍을 이루는 경우에는 100% 함께 나타난 것이기 때문에 PMI 값이 높게 나타나게 된다. 이처럼 논문 전체에 적은 횟수로 출현한 단어쌍이 높은 중요도를 갖는 것을 방지하기 위해, 본 연구에서는 단어간 연관성 $K(x,y)$를 전체 출현 빈도를 고려해 아래와 같이 정의하여 사용하였다.

$$K(x,y) = p(x,y) \times PMI(x,y) = p(x,y) \times \log\frac{p(x,y)}{p(x)p(y)}$$

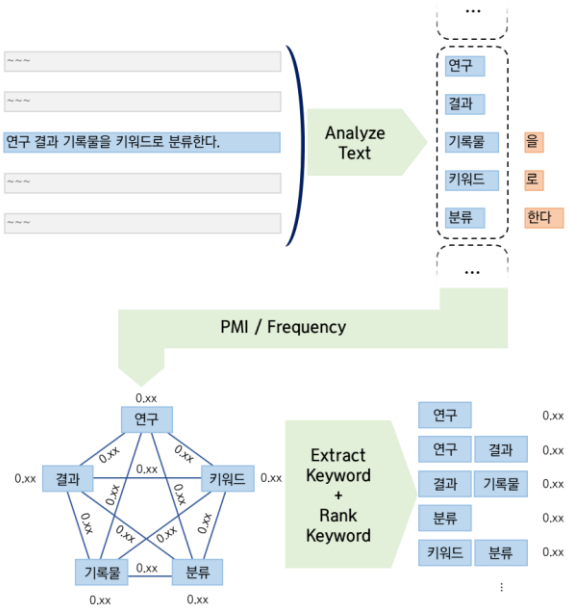
2.3. 연구 검색 시스템 개발 환경

본 시스템에서의 서버는 HP ProLiant ML350p Gen8 Server를 이용하였으며 Ubuntu LTS 16.04 에서 node.js 4.2.6 과 ExpressJS 을 기반으로 한다. 웹페이지의 경우 EJS 2.6.1 과 bootstrap 4.0.0-beta.3 를 바탕으로 제작하였으며, 데이터베이스는 MySQL Server 5.7.23 를 기반으로 구축하였다.

이외에도 다수의 기능 구현을 위하여 multer(파일 업로드), pdf2json(논문 파일 텍스트 추출), python-shell(키워드 추출 알고리즘 구동), body-parser(POST req 처리), express-session(계정 로그인 상태 확인) 등 npm 기반 모듈들을 활용하였다.

3. 설계 및 구현

3.1. 키워드 추출 알고리즘



(그림 1) 키워드 추출 알고리즘 구조도

키워드 추출 알고리즘은 AnalyzeText, PMI/Frequency, ExtractKeword, 그리고 RankKeyword, 총 네 단계로 나눌 수 있다.

AnalyzeText 는 논문의 텍스트를 불러온 후, 문장별로 분리, 형태소 분석을 통해 불필요한 부분들을 모두 제거하는 단계이다. 본 연구에서는 체언, 동사, 형용사를 남기고 이외의 단어들을 제거하였다.

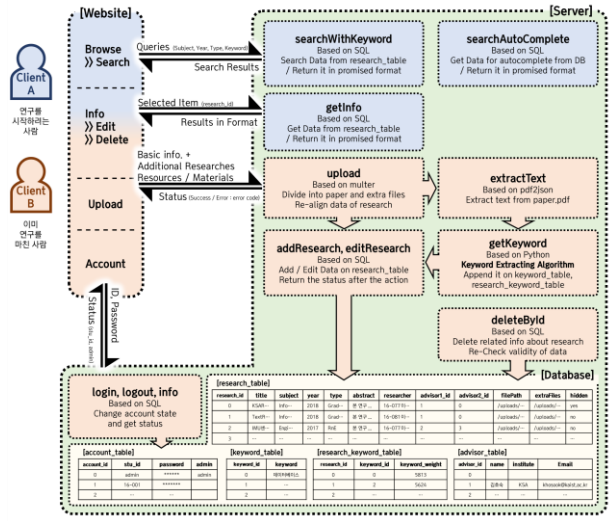
PMI/Frequency 단계에서는 모든 단어에 대해서 (단어 개수)/(문장 개수)로 출현 빈도를 계산한 후, 이를 바탕으로 인접한 모든 단어쌍에 대해서 PMI 값을 계산하였다.

ExtractKeyword 에서는 PMI 값과 빈도수를 바탕으로 각 단어쌍의 K(x,y)를 계산하였고 단어쌍들은 K 값을 가중치로, 한 단어 키워드는 출현 빈도를 가중치로 하여 리스트를 생성한 후, 상위 20 개의 키워드만을 선택하였다.

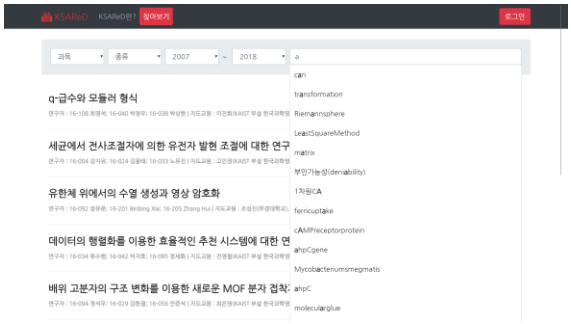
마지막으로, RankKeyword 단계에서는 최종적으로 생성된 키워드 리스트에서 AB 와 BC 등 겹치는 단어가 있는 두 개의 단어쌍을 ABC 의 형태로 합치고, 가중치도 합쳐주어 3 단어 이상의 키워드들을 생성했다. 추가적으로, 제목의 형태소를 분석한 결과와 논문에서 자체적으로 설정한 주제어들의 가중치를 추가로 부여하였다.

[2] https://konlpy-ko.readthedocs.io/ko/v0.4.3/

3.2. 연구 검색 시스템 설계

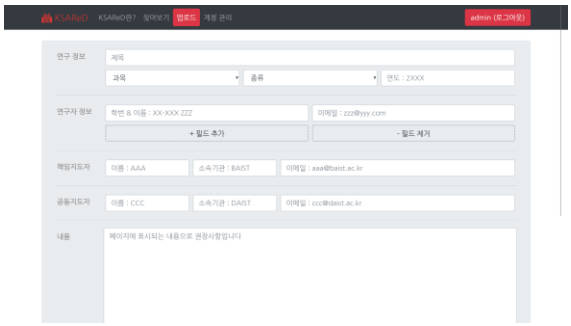


(그림 2) 전체 시스템 구성도



(그림 3) 검색 화면

(그림 4) 연구 정보 화면



(그림 5) 업로드 화면

연구 검색 시스템의 주 사용자는 크게 연구를 시작하려는 사람과 이미 연구를 마친 사람으로 나눌 수 있다.

연구를 시작하려는 사람은 과목, 종류, 연도와 함께 제목, 지도교원, 키워드 등을 추가로 입력하여 선행 연구자료를 검색할 수 있다(searchWithKeyword - 그림 2, 그림 3). 검색 결과에서 연구 선택 시 새 창으로 각 연구의 세부사항을 볼 수 있으며 논문 파일과 연구자가 추가로 업로드한 첨부파일들을 열람할 수 있다(getInfo - 그림 2, 그림 4).

이미 연구를 마친 사람은 로그인하여 업로드 메뉴를 통해 연구 결과를 기록할 수 있으며(upload - 그림 2, 그림 5), 이 때 함께 첨부한 논문 파일은 텍스트로 변환하여 파일 형식으로 저장, 상기 알고리즘을 통해 키워드를 추출하여 데이터베이스에 keyword_table 과 research_keyword_table 으로 함께 저장한다(extractText, getKeyword - 그림 2).

또한 각 연구자의 계정은 account_table 을 통해서 관리되며, 연구 정보, 지도교원의 정보는 각각 research_table 과 advisor_table 을 통해서 관리된다.

4. 구현 결과

4.1. 키워드 추출 알고리즘

체언과 형용사, 동사를 이용하여 키워드 추출 알고리즘을 구현해본 결과, 키워드만을 통해 연구의 대략적인 핵심 요소들을 파악할 수 있을 정도로 키워드가 추출되었다.

예를 들어 “Michelson 간섭계를 이용한 모형 중력파 데이터 분석” 논문 같은 경우에는 ‘중력파’, ‘간섭계’, ‘데이터’, ‘ChirpMass‘, ‘신호’가 키워드로 나왔고, “세균에서 전사조절자에 의한 유전자 발현 조절에 대한 연구”는 ‘유전자’, ‘발현’, ‘ahpC’, ‘Crp’, ‘조건’ 등이 키워드로 나왔다.

하지만 한계점도 분명 존재했다. 첫 번째는 연구에 사용된 형태소 분석기인 Mecab 의 정확성 문제이다. Mecab 은 다른 형태소 분석기에 비해 정확도가 높은 편이지만, 그럼에도 불구하고 ‘참여자’를 ‘참’, ‘여자’로 분류하는 등 잘못된 형태소 분리로 인해 키워드 추출이 의미의 중요성과 무관하게 이루어지는 경우가 존재했다. 특히 과학 계열의 연구논문은 그 특성상 화학식이나 학명 등 일상생활에서는 쓰이지 않는 과학 전문 용어가 많이 포함되어 있다. 이러한 단어들이 대체로 Mecab 분석기에서 완전히 다른 의미를 가진 형태소들로 분할되어서 체언, 동사, 형용사 이외의 형태소로 인식되며 누락되는 등의 이유로 전문 용어가 키워드로 추출이 안 되는 경우도 존재했다.

두 번째는 단어 간 연관성을 하나의 형태소에 대해서 출현 빈도를 가중치로 이용하는 방식 자체의 한계이다. 출현 빈도가 높지만 구체적인 연구의 주제를 파악하는데 크게 도움이 되지 않는 ‘연구’, ‘실험’, ‘방법’등의 일반적인 단어들이 키워드로 추출되었다. 이러한 정확하지 않은 키워드들은 직접 따로

제거하는 방식으로 해결하였다.

마지막으로, 이 연구에서는 한국어 형태소 분석기만을 이용했기 때문에 영어로 쓰인 연구 논문의 경우는 형태소에 대응되는 품사나 그 의미를 분석하지 못하였고, 때문에 자주 쓰이는 'is', 'are' 등의 be 동사와 같은 단어들이 출현 빈도로 인해서 키워드로 추출되는 경우가 많았다. 따라서 이 경우 역시도 일일이 예외처리를 하여 키워드에서 제거하였지만, 영어 문법이 다양한 만큼, 완벽히 제거하지는 못하였다.

4.2. 연구 데이터베이스 웹 서버 구현 결과

웹 서버의 구현 결과 연구자가 본인의 연구 결과물을 쉽게 공유할 수 있는 환경을 갖추었다. 무작위 파일 업로드와 연구 결과를 임의로 수정하거나 삭제하는 등의 문제점을 막기 위해 계정을 만들어 접근 권한에 차이를 두었다.

또한 키워드 생성 알고리즘의 적용으로 키워드를 통한 검색이 가능하여 더 정확하고 직관적인 검색 환경을 만들 수 있었다. "IMU 를 활용한 모듈형 손 모양 인식 장치 개발" 논문은 제목과 주제어에는 없는 '센서'라는 단어로 검색을 해도 찾을 수 있었으며, "키넥트 카메라를 이용한 평면인식 알고리즘 개발" 논문 역시 중요하게 연구에서 큰 비중을 차지하지만 제목과 주제어에는 나타나지 않는 '표면법선벡터'를 검색해서 찾을 수 있었다.

5. 결론 및 향후 연구 계획

본 연구는 논문으로부터 키워드를 추출하여 검색의 효율성에 있어서 제목과 주제어의 한계를 극복하고, 이를 적용하여 연구 검색 시스템을 개발하는 것을 목적으로 한다.

형태소 분석을 통해 불필요한 단어들을 걸러내어 정확도를 높이고 시행시간을 줄였으며, 단어 출현 빈도 및 PMI 지표를 기반으로 정의한 단어간 연관성을 중심으로 하는 키워드 추출 알고리즘을 개발하였다. 해당 분야의 비전문가가 논문을 읽고 키워드를 매기는 것과 대략 비슷한 수준의 결과를 얻을 수 있었으며, 시간이나 노동의 절감 측면에서 큰 이득을 볼 수 있었다. 현재 키워드 추출 알고리즘의 경우, 영문 텍스트 분석이 이루어지지 않고, 화학식이나 학명 등 전문용어 분석이 정확하지 못한 문제점이 있다. 영문 텍스트 분석은 영문용 알고리즘을 별개로 사용하여 해결할 예정이며 전문용어 분석은 전문 용어 사전을 바탕으로 모듈을 보강하여 해결할 계획이다. 또한, 연구 검색 시스템에서는 생성된 키워드가 검색의 정확도와 직관성을 높이는 데에만 사용되고 있으나, 이를 확장하여 키워드의 유사성을 기반으로 현재 검색한 연구와 연관성이 높은 다른 연구를 추천하는 등의 기능을 추가하고자 한다.

본 연구의 결과물은 해가 거듭될수록 새로운 논문들이 계속 추가되면서 최종적으로는 학생들의 연구 성과가 축적된 지식 창고를 만들 수 있을 것이며, 기존의 목표 대상이었던 한국과학영재학교를 넘어서 같은 불편함을 가진 다른 곳에서 더욱 많이 활용될 수 있을 것이라 전망한다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부의 지원을 받아 KAIST 부설 한국과학영재학교의 졸업 연구 프로그램의 일환으로 수행되었습니다.

참고문헌

[1] 신성윤, 이양원. (2010). 비감독 학습 기법에 의한 한국어의 키워드 추출. *한국정보통신학회논문지*, *14*(6), 1403-1408.

[2] 박성건, 김완섭, 이대택. (2017). Mecab-ko 형태소 분석을 이용한 한국체육학회지 연구동향 분석. *한국체육학회지, 56(6)*, 595-605x

[3] 송상일, 이동주, 이상구. (2010). PMI 를 이용한 우리말 어휘의 의미 극성 판단. *한국정보과학회 학술발표논문집*, *37*(1C), 260-265.

[4] Nixon R., (2016). Learning PHP, MySQL & JavaScript With jQuery, CSS & HTML5, 4th ed.