문서의 공기관계를 이용하여 국가 R&D 보고서간 유사도 계산

김남훈*○, 주종민*, 박혁로**, 양형정***, 최광남****
전남대학교 전자컴퓨터공학대학원, 한국과학기술정보연구원
knh3767@naver.com, 3827925@naver.com, hyukro@chonnam.ac.kr, hjyang@jnu.ac.kr, knchoi@kisti.re.kr

Similarity calculation between national R&D reports using co-occurrence

Nam-Hun Kim*○, Jong-Min Joo*, Hyuk-Ro Park**, Hyung-Jeong Yang*** ,Kwang-Nam Choi****
Chonnam National University, Department of Computer Science, Korea Institute of Science and Technology Information

요 약

본 논문에서는 문서의 공기관계를 통해 추출된 문서의 특징을 이용하여 유사 보고서를 판별하는 시스템을 제안한다. 국가 R&D 보고서의 XML형식 파일에서 텍스트를 추출 후, 문장 단위로 나누어 각 문장의 공기관계를 추출한다. 그 후 공기관계의 노드와 엣지를 문서에 추가하고, 노드로 사용된 단어만 남기고 나머지 단어는 제외한다. 그리고 이것을 문서의 특징으로 삼고 유사도 계산을 한다. 이 때, 유사도 계산은 코사인 유사도를 사용한다. 실험결과, 국가 R&D문서 유사도 계산에서 제안된 방법이 기존의 방법보다 높은 분류율을 보여주었다.

주제어: 국가 R&D 보고서, 공기관계, 유사도계산, 자연어처리

1. 서론

각 연구기관과 국가에서는 R&D 투자의 확대와 투자의 효율성을 높이기 위하여 연구사업 선정과정에서 중복과제와 유사과제를 검토하는 과정을 거친다. 현재는 검색엔진의 검색결과에 의존하여 유사과제를 파악하고 있다. 그러나 이 방식은 기관에 국한된 자료에 근거하거나, 키워드 매칭 검색결과의 단점인 제한되거나 너무 광범위한 범위의 정보를 기반으로 판단하므로 유사과제 파악이 어렵다.

따라서 국가가 지원하는 연구 개발지원 사업에서 유사한 과제를 중복 제안하는 것을 사전에 방지하기 위해 유사문서를 효율적으로 판별해 내는 알고리즘이 필요하다.

본 논문은 현재 사용 중인 국가과학기술종합정보서비스(NTIS)의 시스템 중의 하나인 유사과제 검색 시스템에 필요한 알고리즘을 개선하는데 그 목적이 있다. 본 연구에서는 문서의 색인어들에 대한 가중치를 부여하는 벡터 공간검색(Vector-Space Retrieval)모델의 한 종류인 TFIDF(Term Frequency Inverse Document Frequency)를 기본 구조로 채택하고, 유사도 계산은 코사인유사도(Cosine similarity)를 이용하였다.

본 연구는 서론에 이어서 2장에서는 관련연구, 3장에서는 알고리즘 개발, 4장에서는 실험 및 알고리즘 평가를 다루고, 마지막으로 5장에서는 결론 및 연구의 한계점을 기술한다.

2. 관련 연구

유사 과제를 식별하기 위한 시스템은 현재 상당수 개

* 전남대학교 전자컴퓨터공학과 석사과정
** 교신저자, 전남대학교 전자컴퓨터 공학과 교수
*** 전남대학교 전자컴퓨터 공학과 교수
**** 한국과학기술정보연구원, NTIS센터

1.본 연구는 한국과학기술정보연구원에서 수행하는 (국가 R&D정보의 공유/협력 강화로 국가과학기술 가치 극대화) 사업의 위탁연구로 수행되었습니다.
2.This Research has performed as a subproject of project No (Maximize the Value of National Science and Technology by the KOREA INSTITUTE of SCIENCE and TECHNOLOGY INFORMATION (KISTI)

발되어 있다. 기본적인 유사도 측정은 두 문서간의 나오는 색인어를 비교분석하여 계산한다.

논문 [1]에서는 기본적으로 위와 같이 문서 간에 나오는 색인어를 통해 유사도를 계산하는데 이것을 발전시켜서 과제의 문서를 분석하여 색인어들을 추출하고 각각에 가중치를 부여한다.

그리고 TFIDF기법을 적용하여 기존 문서들과 비교하여 유사문서를 찾아낸다. 또한 검색속도 향상을 위하여 K-최근접 문서(KNN:K-Nearest Neighbors) 기법을 적용한다.

논문 [2]에서는 색인어 추출 단계에서 Document Vector를 기반으로 한 검색엔진에 연구보고서 초록을 추가한다. 그리고 분석단계에서 과제 키워드에서 복합 키워드 중심으로 생성한 과제의 연구 주제망과 항목별 가중치를 활용하여 유사도를 측정한다. 항목별 가중치는 과제명(5), 연구책임자(1), 연구목적(3), 연구내용(3), 기대효과(3), 키워드(1), 연구보고서 초록(3)으로 적용한다. 실험결과 연구보고서 초록이 유사도에 영향을 미치고 있고, 복합 키워드 기반의 연구 주제망과 항목별 가중치를 활용하였을 때 유사도에 대한 정확도를 판단할 수 있는 범위가 확대되는 것을 확인하였다.

3. 실험환경 및 알고리즘 개발

3.1 실험환경

본 연구는 KISTI(Korea Institute of Science and Technology Information)에서 제공한 100건의 무작위 연구보고서를 데이터베이스로 활용하였다.

그림 1 제공받은 XML문서 구조

각 문서는 XML형식으로 구조화 되어있고, XML구조는 통일 되어 있지 않다. 포함된 문서는 분야가 나누어져 있지 않고, 서로 다른 문서이다. 현재 유사한 문서라고 표현할 만한 정답문서가 존재하지 않아, 전체 문서에서 저자가 직접 유사 문서를 찾아 두개의 서로 유사한 문서를 각각 서론, 본론, 결론파트로 나누어서 두 문서를 조합하여서 총 8개의 유사 문서를 만들어 내었다. 이렇게 해서 3 종류의 유사문서 정답집합이 존재한다. 그리고 이것이 유사도 계산 알고리즘의 판단 기준이 될 것이다.

소프트웨어 환경은 다음과 같다. XML문서에서 자바를 이용하여 텍스트를 추출하고, 문장 단위로 분할을 한다. 그 후 R에서 불용어 처리를 하고, 형태소 분석을 거친 다음 각 문서의 공기관계를 한 문장을 기준으로 추출 하였다. 그 다음 자바를 이용하여 이 공기관계를 문서의 특징으로 삼고, 공기관계 만큼 가중치를 주고, 또한, 공기관계에 나온 단어만을 남기고 나머지 단어는 다 제외하였다. 그리고 이 데이터를 가지고 자바 라이브러리인 루씬을 사용하여 각각 색인 후 유사도를 계산 하였다.

3.2 알고리즘의 개발

알고리즘은 3단계로 나눌 수 있다. 첫째는 색인, 둘째는 공기관계 추출과 가중치 적용, 마지막으로 유사도를 계산하는 것이다.

3.2.1 색인

본 단계에서는 한글 형태소 분석 R 패키지인 Konlp를 사용하여 조사와 부사, 접속사, 그리고 의미를 갖지 않는 기호를 제거한다. 그리고 형태소 분석 전에 불용어를 먼저 제거한다.

일반적인 한글 불용어 사전 외에도 연구제안서에 맞는 불용어 리스트가 필요하여, KISTI에서 제공해준 100개의 연구보고서에서 형태소 분석 후에 나온 어휘들 중 빈도수가 10이상인 고빈도 어휘를 불용어 처리 기준으로 설정하였다. “연구, 결과, 수행, 사용, 개발, 방법, 분석” 위의 단어는 R&D 연구보고서에서 너무나 많이 사용되는 단어이면서 동시에 문서의 특징으로 삼기에는 가치가 없다고 판단되어 불용어로 처리하였다.

3.2.2 공기관계 추출과 가중치 적용

형태소 분석과 불용어를 처리한 후, 한 문장을 기준으로 하여 공기관계를 추출하였다. 공기관계란 한 문장 안에서 동시에 출현한 단어를 의미하는데, 한 문장 안에서 같이 출현했다는 것은, 그 단어끼리 서로 의미적으로 연관이 있다고 판단하여 그것을 문서의 특징으로 정하였다.

공기관계로 추출된 어휘들은 네트워크로 표현 될 수

있다. 노드는 추출된 단어이고, 각 노드가 엣지로 연결된 것은 각 단어가 한 문장 안에서 같이 출현했다는 의미이다. 그래서 문서의 특징이라고 생각하는 데이터를 생성할 때, 공기관계뿐만 아니라, 단어들을 공기관계로 추출된 횟수만큼 가중치를 주었고, 또한 공기관계로 추출되지 않은 단어들은 데이터에서 제외시켰다.

3.2.3 유사도계산

이 단계에서는 질의문서와 저장된 질의문서와 유사한 문서 그리고 유사 하지 않은 나머지 R&D보고서를 유사도 계산하여 보여준다.

유사도계산은 자바 라이브러리인 루씬을 사용하였고, 코사인 유사도를 이용하여 유사도를 계산한 다음 유사도가 가장 높은 것부터 순서대로 보여준다.

4. 실험 및 알고리즘 평가

4.1 전체 시스템 구조도

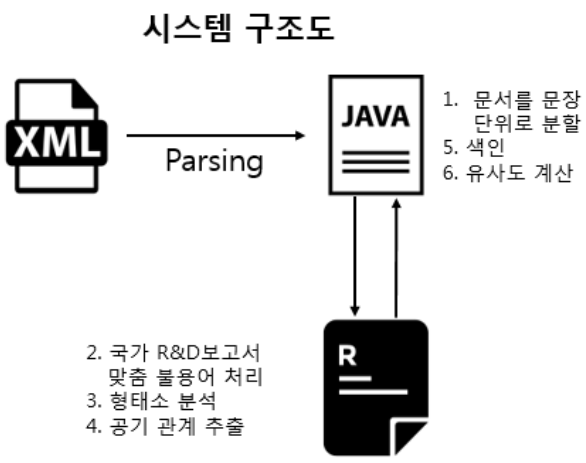

그림 2 제안된 알고리즘의 전체 시스템 구조도

4.2 실험과 알고리즘 평가

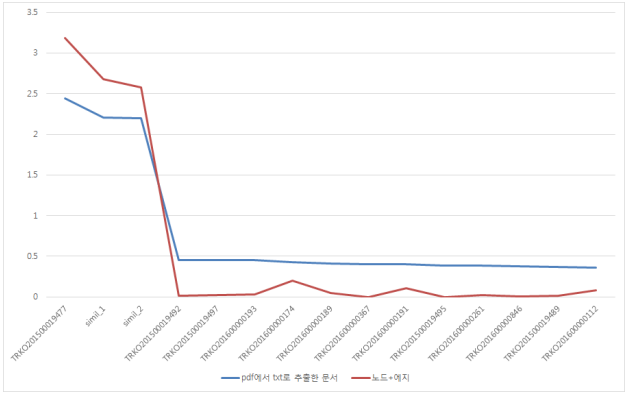

그림 3 유사보고서 TRKO2015000019477에 대한 유사도 값

그림 3는 유사보고서 TRKO2015000019477에 대한 다른 87개의 보고서를 대상으로 유사도 값을 구한 결과에서 상위 15개만을 나타내었다. 서로 유사한 보고서인 simil_1, simil_2는 유사도 값이 높게 나온 것을 볼 수 있다. 파란색 선은 기존의 방법인 단순 색인어만을 이용하여 유사도 값을 계산한 것이고 빨간색 선은 제안하는 방법인 공기관계를 이용한 유사도 값이다. 실험 결과 제안하는 방법이 유사한 보고서끼리는 더 높은 유사도 값을 내고, 유사하지 않은 보고서는 더 낮은 유사도 값을 가진 것을 확인하였다.

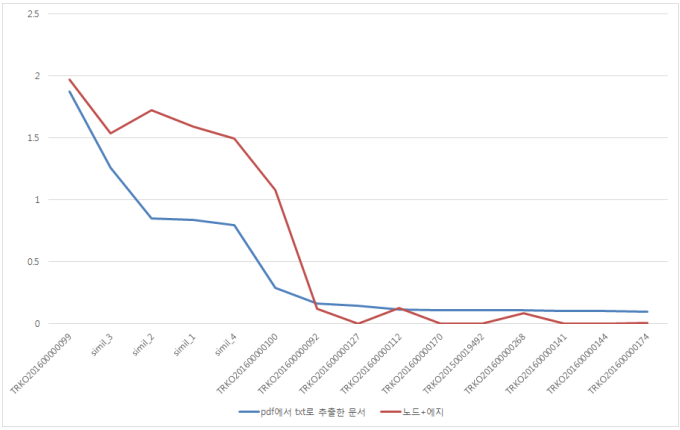

그림4 유사보고서 TRKO2016000000099에 대한 유사도값

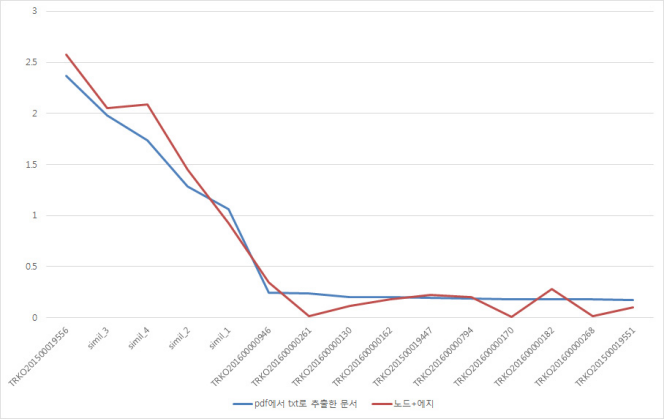

그림 5 유사보고서 TRKO2015000019556에 대한 유사도 값

하지만 모든 유사보고서에 대해서 좋은 결과를 가져오지는 못했다. 그림 4의 유사보고서 TRKO2016000000099에 대한 유사도 값에서 몇몇 문서에 대해서는 기존의 방식과 제안한 방식이 서로 비슷한 유사도 값을 가지는 경우도 있었다. 그림 5의 유사보고서 TRKO2015000019556에서는 서로 유사한 문서인 simil_3에서 오히려 기존의 방식이 더 높은 유사도 값을 가졌고 유사하지 않은 문서에 대해서는 제안하는 방식이 더 높은 유사도 값을 가져서 분별을 하기가 더 어려워졌다.

전체적으로 성능이 향상되기는 하였지만 몇몇 문서에 대해서는 오히려 성능이 떨어지는 현상도 있어서 불용어 처리와, 색인어 추출 등의 개선이 필요하다.

5. 결론 및 연구의 한계점

본 연구를 통하여 개발된 알고리즘은 유사과제를 판별함에 있어서 기존의 방법보다 개선된 결과를 가져왔다. 유사한 문서의 경우 유사도 값이 증가하였고, 유사하지 않은 문서의 경우는 유사도 값이 감소했다. 공기관계를 이용함으로써 기존의 방법보다 유사문서 분별 능력이 향상되었다는 것을 알 수 있다. 본 알고리즘은 유사과제 판단을 위한 것이지만, 향후에는 국가 R&D보고서 뿐 만 아니라 다른 분야에도 적용이 가능 할 것으로 보인다. 본 알고리즘은 구조와 방법은 단순한데, 그에 반하여 성능은 뛰어남을 알 수 있다.

본 연구는 유사과제 분별을 위한 시스템에서 아직 더 많은 고려가 필요하다. 특히 불용어 사전과 동의어, 유사어 뿐 만 아니라, 국제화 시대인 만큼 영어에 대한 부분도 아직 미흡한 부분이 있다. 또한 개발된 알고리즘에 대한 복잡도 및 소요시간 등에 대한 객관적인 분석이 전혀 이루어지지 않아서 실제 적용을 위해서는 위의 부분에 대한 체계적인 분석이 필요하다.

참고문헌

[1] 박동진, 최기석, 이명선, 이상태, “유사과제파악을 위한 검색 알고리즘의 개발에 관한 연구”, 한국콘텐츠학회논문지, 제9권, 제11호, pp.54-62, 2009.

[2] 정옥남, 류성열, 김종배, “과제 유사도 측정 개선 모형에 관한 실증적 연구”, 한국디지털콘텐츠학회논문지, 제12권, 제4호, pp.457-465, 2011.

[3] 류창건, 김형준, 조환규, “한국 말뭉치를 이용한 한글 표절 탐색 모델 개발”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제14권, 제2호, pp.231-235, 2008.

[4] 송광호, 민지홍, 이가영, 김유성, “유의어 사전을 이용한 문서 내 표절 구간 탐색 시스템 개발”, 한국정보과학회 제 41회 정기총회 및 동계학술발표회, pp.1385-1387, 2014.

[5] 박선영, 김지훈, 김선영, 김형준, 조환규, “대용량 문서 집합에서 유사 문서 탐색을 위한 효과적인 전처리 시스템의 설계”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제16권, 제5호, pp.626-630, 2010.