연구 보고서의 공기관계 정보에 제목 및 요약의 가중치를 적용한 유사도 계산

김남훈† , 주종민† , 박혁로† 1), 양형정†
† 전남대학교 전자컴퓨터공학대학원

Calculation of similarity by weighting title and summary in word co-occurrence of research reports

Nam-Hun Kim† , Jong-Min Joo† , Hyuk-Ro Park† Hyung-Jeong Yang†
† Chonnam National University, Department of Computer Science

요 약

본 논문에서는 국가 연구 보고서의 공기 관계 정보와 제목, 요약 등에 가중치를 적용한 유사도 계산 방법을 제안한다. 이를 위해 국가 연구개발 보고서에서 텍스트를 추출하여 한 문장 단위로 문서를 분할하고, 기본 불용어와 보고서에서 특징적으로 나타나는 불용어를 처리하고 형태소 분석을 한 뒤 공기관계를 추출하였다. 또한 문서의 유사도 계산시 정확성을 높이기 위해 제목과 요약 부분에 가중치를 부여하였다. 이를 통해 본 논문에서 제안하는 방법이 문서 검색 라이브러인 루씬(Lucene)을 이용한 방법보다 2.5%의 검색성능 향상을 그리고 Knn-휴리스틱 방법보다는 1.1%의 검색성능 향상을 보였다. 이러한 결과를 통해 문서의 요약과 제목 그리고 공기관계 정보가 연구보고서의 유사도를 계산 하는데 영향을 미친다는 것을 보였다.

1. 서 론

각 연구기관과 국가에서는 연구개발 투자의 확대와 투자의 효율성을 높이기 위하여 연구사업 선정과정에서 중복 및 유사 과제를 검토하는 과정을 거친다. 현재는 키워드 매칭 기반의 검색엔진 검색결과에 의존하여 유사과제를 파악하고 있다. 그러나 이 방식은 키워드 매칭 검색결과의 단점인 특정 단어로 제한되거나 너무 광범위한 범위의 정보가 검색될 수 있어 유사 문서의 판단에서 성능을 제한하는 요인이 된다. 따라서 국가가 지원하는 연구 개발 지원 사업에서 유사 연구 보고서의 특징을 고려한 유사도 계산 알고리즘이 필요하다.[6][16]

[1]의 연구는 한글의 형태소적인 특징을 이용하여 문서를 축약하는 기술을 적용하여 문서 표절 검사의 성능을 개선 하였다. [2]의 연구는 기사에서 구문을 추출하고 그 구문들을 질의문서로 웹 검색을 실시하여 결과 값을 이용하여 해당문서의 표절여부를 검사하는 연구를 하였다. 하지만 위의 연구는 정확한 중복문장을 판별해야하는 표절에 관한 연구이므로 의미적인 유사성을 판별하는 유사문서 검색에 적용하기는 어려움이 있다. [3]의 연구는 형태소 분석을 통해 추출한 용어 중에서 특별한 용어인 주제어 용어를 선별하는 방법을 이용하여 문서 유사도를 개선하는 방법을 연구하였다. 이 연구는 문서의 유사성을 검사하기 위해 주제어와 관련된 용어의 개수를 바꿔가며 실험 했으나 기본적으로 형태소 분석 기술을 이용한 용어 중심의 방법을 적용한 것이다. 형태소 해석 기술을 이용한 용어 중심의 방법들은 형태소는 언어마다의 고유 특성으로 인해 단어 의미에 애매한 부분들을 가지고 있으므로, 문서가 가지고 있는 의미를 제대로 표현하는데 한계가 있다.

공기관계를 이용하는 [4]는 키워드 자동추출 기법에 높은 성능을 보였다. 단어와 단어사이의 공기 관계는 동시 출현하는 단어를 고려함으로써 유사성을 판별할 때 성능을 높여준다[13]. [5][6]은 연구보고서의 유사성 판별 할 때 제목과 요약 부분에 가중치를 주는 것이 유사 문서 측정의 정확도를 높여준다는 것을 검증하였다. 즉, 연구 보고서라는 특성 상 문서의 요약과 제목은 문서의 의미와 내용을 함축적으로 나타내기 위해 작성한 것으로 문서의 유사도 판단에 기여함을 보이고 있다.

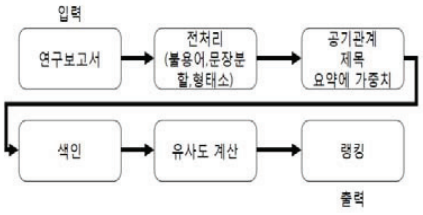
본 논문에서는 연구보고서의 문서 유사도 측정의 성능을 높이기 위해 단어사이의 동시 출현 빈도인 공기관계와 제목과 요약 부분에 가중치를 부여하는 방법을 제안한다[17]. 이를 위해 먼저 연구 보고서에서 단어사이의 공기관계를 추출하고, 제목과 요약 부분에 가중치를 부여하고[5] 문서사이의 유사도를 측정한다.

1) 교신저자

성능 비교를 위해 루씬(Lucene)에서 추출한 문서 랭킹으로 11-포인트 평균 정확률(11-point average precision)을 적용하여 정확률과 재현률의 평균값을 비교한 결과 2.5%의 성능 우수를 보였다.

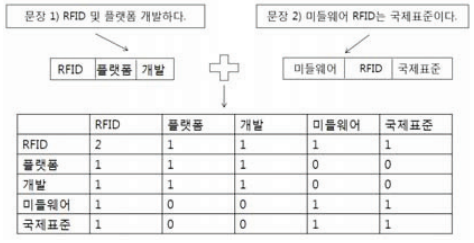
2. 본 론

앞서 언급한 바와 같이 본 논문에서는 공기관계와 제목 및 요약의 가중치를 적용한 새로운 유사도계산 방법을 제안하고자 한다. 기존의 단어의 빈도만을 이용한 색인 및 유사도 계산 방법은 문서의 의미적인 내용을 반영하지 못하고 출현 빈도가 높은 단어가 문서의 내용을 대표하게 된다. 하지만 이것은 연구보고서의 경우 같은 주제로 연구한다고 해도 세부적인 내용에 다른 부분이 많기 때문에 유사문서 판단에 부정확성이 발생한다. 본 논문에서는 유사문서 판단에 공기관계 및 제목과 요약의 가중치를 반영한 방법을 제안한다.



[그림 1] 시스템 구조도

이 기법은 그림 1과 같이 총 6단계의 과정으로 이루어진다. 첫 번째는 전처리 단계로 문서를 문장단위로 분리하고 명사들을 추출하여 저장한다. 이때 그 단어가 복합명사일 경우 복합명사를 구성하는 구성명사도 함께 출현하는 것으로 처리한다.



	RFID	플랫폼	개발	미들웨어	국제표준
RFID	2	1	1	1	1
플랫폼	1	1	1	0	0
개발	1	1	1	0	0
미들웨어	1	0	0	1	1
국제표준	1	0	0	1	1

[그림 2] 연구보고서의 동시출현 단어 행렬 생성 예

두 번째 단계인 공기관계와 제목 및 요약에 가중치를 곱하는 단계는 각 문장에서 출현한 단어 쌍들의 출현횟수를 저장한 공기관계 행렬을 생성한다. 추출한 공기관계 단어 행렬은 그림 2와 같다. 이 행렬의 대각성분 즉, 동일한 단어가 교차하는 성분은 문서 전체에서 그 단어가 출현한 개수이고 서로 다른 두 단어가 교차하는 성분은 그 두 단어가 함께 나온 문장의 개수이다.

$$cscore(D_1, D_2) = \sum_{j \in \{공기관계, 제목, 요약\}} c_j * score(D_{1j}, D_{2j})$$
$$c_j = weight\ constant\ of\ each\ part \quad (1)$$
$$D = (a_{j1}, a_{j2}, \ldots..a_{jn}), j \in \{공기관계, 제목, 요약\}$$

유사도 계산식 $cscore$은 다음 식 1[11]과 같다.

식 1에서 사용된 c_j의 값은 공기관계의 가중치다. 제목 과 요약 부분을 따로 추출 해 저장 한 후 식 1의 값을 이용하여 공기관계, 제목, 요약의 가중치가 적용된 유사도 값을 구하고, 루씬[12]을 이용하여 문서 유사도 값을 구한 후 식1의 값과 루씬의 유사도 값을 합하여 최종 유사도 값에 따른 랭킹이 산출된다.

3. 실험 및 결과

3.1 데이터 집합

본 논문에서는 NDSL(National digital Science Library)에 업로드 된 국가 R&D 보고서 가운데 6개 주제를 선정하여 실험에 이용하였다. 선정된 주제는 RFID, USB, 게임, 모바일, 영상이다. 전체 실험 문서 120개는 같은 주제별로 분류 된 문서이다. 문서를 구분하는 기준은 NDSL에서 전문가가 같은 주제로 분류해놓은 것을 본 실험에서도 같은 기준으로 이용 하였다. 각 문서는 평균 1592개 정도의 단어를 포함한다.

3.2 실험방법

실험을 위해 각 연구 보고서는 문장 단위로 분리하고 기본적인 한글 불용어 사전 외에 연구보고서 분야에서 필요한 불용어 리스트를 구성한다. 형태소 분석 후 상위 15%에 해당되는 고빈도 단어들을 불용어 처리 기준으로 설정하였다. 예를 들면, “연구, 개발, 내용, 실험, 분석, 제시, 시스템 ...” 등과 같은 단어들은 연구 보고서에 빈번하게 많이 사용되는 단어이므로 식별력이 없기 때문에 이들을 불용어로 처리한다[7].

공기관계의 출현율을 0.5부터 0.1까지 0.1단위로 실험하고, 0.1부터 0.01까지 0.01 단위로 실험한 결과 같은 분야의 문서가 가장 상위에 랭크되는 값인 0.02로 고정하여 실험을 진행하였다. 제목과 요약 부분의 가중치는 [5]의 연구에 따라 제목은 5, 요약은 3의 값이 식 1의 c_j 값으로 적용 되었다.

형태소분석, 불용어, 공기관계, 가중치를 처리한 후 성능 평가를 위해 고성능 정보 검색 라이브러리인 루씬[12]을 이용하여 유사도 계산 실험을 진행하였다. 정보검색 시스템 성능 평가에는 카테고리 랭킹 시스템, 이진 분류 시스템[14], F-점수가 사용되고 있다. 본 논문에서는 연구보고서의 경우 주제별로 분류가 되어 있

어 카테고리 랭킹 시스템이 가장 알맞기 때문에 카테고리 랭킹 평가 방법인 11-포인트 평균 정확률로 성능을 측정 하였다.

11-포인트 평균 정확률[14]은 전체 테스트 문서 집합에 대한 전체적인 평가를 하기 위해 각 문서별로 재현율에 따른 정확률을 측정한 뒤 전체 문서 집합에 대해 평균을 내어 산출하는 방식이다. 정확률은 다음 식 2과 같이 표현된다.

$$\text{정확률} = \frac{\text{시스템에 의해 판단된 적합 문서수}}{\text{시스템이 적합하다고 판단한 문서수}} \quad (2)$$

재현율은 다음 식 3와 같이 표현된다.

$$\text{재현율} = \frac{\text{시스템에 의해 판단된 적합 문서수}}{\text{적합 문서수}} \quad (3)$$

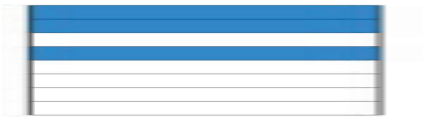
11-포인트 평균 정확률은 식 4와 같이 단계의 재현율을 가지고 각 단계마다 정확률을 측정하고 평균화한다. 그리고 각 단계는 0.0부터 1.0 까지 0.1 단위로 나누어진다.

$$AP = \frac{\sum_{k=1}^{|S|} P(k)}{|S|} \quad (4)$$

AP=각 재현율에서의 정확률을 계산하고 평균을 낸 값

그림 3은 3번째까지의 재현율을 측정 했을 경우 2개는 정답이고 1개는 오답을 의미한다. 즉, 그림 3로 AP를 계산한 것은 아래와 같다.

$$\frac{P@1 + P@2 + P@4}{3} = \frac{1/1 + 2/2 + 3/4}{3} = 0.92$$



[그림 3] AP결과 예시

3.3 실험결과

표 1에 나타난 Knn-휴리스틱 방법[6]은 제목, 저자, 소속단체, 요약에 각각 다른 가중치를 주고, 색인어를 선택 할 때 각 질의 문서의 색인어들 중에서 가중치 값을 상위부터 K개 결정한다. 루씬[12]는 문서 검색 라이브러리인 루씬에 연구 보고서를 입력 값으로 넣어 추출된 랭킹으로 평가한 결과이다.

<표 1> 제안된 방법의 평균 정확률 비교

방법	평균 정확률(%)
제안한 방법	**65.6%**
Knn-휴리스틱[6]	64.5%
루씬[12]	63.1%

본 논문에서 제안한 방법을 적용 했을 경우 루씬 검색 방법과 비교하여 표 1과 같이 2.5%의 성능 향상을 보이고, Knn-휴리스틱 방법과 비교하여 1.1%의 성능 향상을 보였다[15]. 재현율에 따른 평균 정확률은 데이터 중 유사 문서 집합을 질의문서로 루씬에 입력하여 추출되는 랭킹을 이용하여 구한 각 주제별 평균값이다. 표 1은 각 분야에서 추출된 재현율에 따른 평균 정확률 값을 총 합하여 다시 11로 나누어 평균 정확률 값을 구하였다.

<표 2> 재현율에 따른 평균 정확률의 변화 비교

재현율	루씬	knn-휴리스틱	**공기관계+제목+요약 가중치적용**
0.0	1.0000	1.0000	**1.0000**
0.1	0.9444	0.9612	**0.9778**
0.2	0.7324	0.8562	**0.8762**
0.3	0.7010	0.7661	**0.7768**
0.4	0.6784	0.7005	**0.6952**
0.5	0.6173	0.6061	**0.6149**
0.6	0.5747	0.5486	**0.5739**
0.7	0.5198	0.4972	**0.5017**
0.8	0.4688	0.4429	**0.4577**
0.9	0.3887	0.4064	**0.4171**
1.0	0.3172	0.3174	**0.3282**

표 2에 각 방법에 대한 재현율에 따른 평균 정확률 값을 보이고 있다. 전체적으로 본 논문에서 제시하는 방법이 더 높은 평균 정확률을 가진다는 것을 알 수 있다.

4. 결론

본 논문에서는 공기관계와 제목 그리고 요약 부분의 가중치를 적용하여 국가 R&D 보고서 유사도 계산에 적용하는 방법을 제안하였다. 실험 결과를 통해 루씬을 적용한 국가 R&D 보고서 대상 유사도 계산 방법에 비해 공기관계와 제목 그리고 요약에 가중치를 주는 방법이 더 높은 성능을 보였다.

향후 연구로는 본 논문에서 제안한 방법이 자칫

한 분야에 치중되어 편향된 평균 정확도가 나올 가능성이 있으므로 문서의 분야를 가리지 않고 문서 검색에 적용할 수 있는 방법을 연구하고자 한다. 또한 개발된 알고리즘에 대한 복잡도에 대한 체계적인 분석이 필요하다.

Acknowledgement

본 연구는 한국과학기술정보연구원(KISTI)의 위탁연구 과제로 수행한 것입니다.

참 고 문 헌

[1] 전명재, 박상돈, 박웅, 허진영, & 조환규. (2004). 한글 구조특성과 지역정렬 알고리즘을 사용한 표절 판정 시스템의 개발. 한국정보과학회 학술발표논문집, 31(2 I), 727-729.

[2] 조동욱, 홍윤선, & 조선옥. (2003). 효과적인 e-런닝 시스템 구축을 위한 과제물 표절 검사. 한국콘텐츠학회 종합학술대회 논문집, 1(2), 53-59.

[3] 장성호, & 강승식. (2003). 용어 선별 기법에 의한 유사 문서 판별 시스템. 한국정보과학회 학술발표논문집, 30(1B), 534-536.

[4] 송광호, 민지홍, & 김유성.(2016). 한글 문서의 단어 동시 출현 정보에 개선된 TextRank를 적용한 키워드 자동 추출 기법. 한국정보과학회 학술발표논문집,28(2B) 64-65,

[5] 정옥남, 류성열, & 김종배. (2011). 과제 유사도 측정 개선모형에 관한 실증적 연구. 한국디지털콘텐츠학회 논문지, 12(4), 457-465.

[6] 박동진, 최기석, 이명선, & 이상태. (2009). 유사과제파악을 위한 검색 알고리즘의 개발에 관한 연구. 한국콘텐츠학회논문지, 9(11), 54-62.

[7] 류창건, 김형준, & 조환규. (2008). 한글 말뭉치를 이용한 한글 표절 탐색 모델 개발. 정보과학회논문지: 컴퓨팅의 실제 및 레터, 14(2), 231-235.

[8] 박선영, 김지훈, 김선영, 김형준, & 조환규. (2009). 대용량 문서 집합에서 유사 문서 탐색을 위한 효과적인 전처리 시스템의 설계. 한국정보과학회 학술발표논문집, 36(2A), 76-77.

[9] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

[10] 김선. (2001). 유전 알고리즘을 이용한 웹 문서 검색 (Doctoral dissertation, 서울대학교 대학원).

[11] 강원석, & 황도삼. (2014). 구문의미분석을 이용한 유사문서 판별기. 한국콘텐츠학회논문지, 14(3), 40-51.

[12] 김동민, 최진우, & 우종우. (2014). 루씬을 이용한 빅데이터 인덱싱 및 검색시스템의 설계 및 구현. 인터넷정보학회논문지, 15(6), 107-115.

[13] 박호진, & 김재훈. (2002). 공기정보를 이용한 한국어 요약 시스템의 성능개선.

[14] Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In Icml (Vol. 97, pp. 412-420).

[15] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98, 137-142.

[16] Buttler, D. (2004). A short survey of document structure similarity algorithms (No. UCRL-CONF-202728). Lawrence Livermore National Laboratory (LLNL), Livermore, CA.

[17] Debole, F., & Sebastiani, F. (2003, March). Supervised term weighting for automated text categorization. In Proceedings of the 2003 ACM symposium on Applied computing (pp. 784-788). ACM.