Journal of Digital Contents Society Vol. 12 No. 4 Dec. 2011(pp.457 - 465)

과제 유사도 측정 개선모형에 관한 실증적 연구

정옥남*, 류성열**, 김종배***

요 약

지난 5년간 우리나라 R&D투자는 연평균 12.2%씩 증가하고 있다. 연구개발 중복 투자 방지와 독창성 도출을 위해서는 유사·중복과제 수행의 사전방지가 필요하고, 이를 위해 과제 유사도의 정확도를 개선할 필요가 있다. 본 논문에서는 유사·중복과제 수행의 사전방지를 위한 과제 유사도 측정 개선모형을 제안한다. 과제 유사도 측정 개선모형은 크게 두 단계로 정의된다. 먼저 추출단계에서 Document Vector를 기반으로 한 검색엔진에 연구보고서 초록을 추가한다. 다음은 분석단계에서 과제 키워드에서 복합 키워드 중심으로 생성한 과제의 연구주제망과 항목별 가중치를 활용하여 유사도를 측정한다. 실험결과 과제정보만을 활용한 기존방식보다 연구보고서 초록을 활용한 개선모형의 유사도가 평균 0.19이상 개선되었고, 단순키워드를 활용한 기존방식보다 복합 키워드 기반의 연구주제망과 항목별 가중치를 활용한 개선모형의 유사도가 평균 9.25이상 감소되었다. 연구보고서 초록이 유사도에 영향을 미치고 있고, 복합 키워드 기반의 연구주제망을 활용함으로써 유사도에 대한 정확도를 판단할 수 있는 범위가 확대되는 것을 확인하였다. 또한, 추가된 사항의 폭이 넓으면 넓을수록 유사도의 정확도가 높아지는 것과 과제정보 등 검색대상의 모집단이 클수록 과제 유사도의 정확도가 높아지는 것도 실험을 통해 확인하였다.

An Empirical Study on Improvement model for Measuring of Project Similarity

Ok-Nam Jung*, Sung-Yul Rhew**, Jong-bae Kim***

Abstract

The annual R&D investment in Korea increased by an average of 12.2percent during the last 5 years. Therefore, prevention of duplicate projects being performed became an important factor in promoting the efficiency of R&D investment and the originality of R&D projects. On measuring the similarity of projects, the measurement model used to estimate the accuracy of the similarity is crucial. In this paper, we propose an advanced measurement model on checking the similarity of R&D projects for promoting the efficiency of R&D investment. The proposed model is made up of the following steps for the model measurement, sampling and analyzing. During the sampling step, we append the abstract of R&D reports on the search engine based on document vector. We then measure the similarity on projects to use research title network which ~~is~~ consists of the compound keyword and the weight of items ~~on~~ during the analysis. The proposed method improved the accuracy for measuring the similarity of projects by an average of 0.19 over the existing search engine and by 9.25 over the simple keyword search on R&D projects. On searching the similarity with the appending conditions and high sampling, it improved the accuracy of measuring the similarity of R&D projects.

Keywords : Similarity, Complex Keyword, U-WIN, Document Vector

1. 서론

지난 5년간 우리나라 연구개발투자는 정부의 적극적인 과학기술정책에 힘입어 타 분야에 비해 큰 폭으로 증가하는 추세이다. 연구개발 투자 증대가

※ 제일저자(First Author) : 정옥남
접수일:2011년 11월 08일, 수정일:2011년 11월 26일
완료일:2011년 12월 06일
* 국가과학기술위원회
jon77@nstc.go.kr
** 숭실대학교 컴퓨터학부
*** 숭실대학교 IT정책경영학과(교신저자)

국가과학기술력 강화로 나타나기 위해서는 유사·중복과제 수행을 사전에 방지하여 중복투자를 최소화할 필요가 있다.

현재, 정부에서는 국가연구개발 과제 기획 및 선정 시 동일하거나 유사한 연구과제를 정부 각 부처에 중복 제안하여 지원 받는 사례의 사전 방지를 위해 국가연구개발사업관리 등에 관한 규정에 의거 과제 협약 전 국가과학기술지식정보서비스(NTIS, National Science & Technology Information Service, 이하 NTIS)를 통한 과제 유사성 검토를 의무화하고 있다. 국가과학기술위원회가 운영하고 있는 NTIS는 국가R&D 사업, 인력, 연구장비, 성과 등 국가가 진행하는 R&D사업 관련 정보를 15개 부처·청과 실시간으로 연계하여 제공하는 시스템으로 매년 국가연구개발사업 조사·분석을 통해 수집되는 30개 R&D부처의 과제정보를 바탕으로 신규 기획 과제에 대한 유사도를 측정 해주고 있다. 과제의 유사도를 측정하기 위해서 과제정보로부터 추출하는 항목은 과제명, 연구목표, 연구내용, 기대효과, 연구책임자, 키워드 등 총 6개이다.

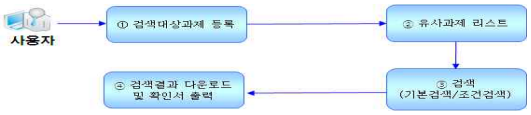
신규로 기획되고 있는 과제가 기존에 수행된 과제와 얼마나 유사한가에 대한 유사도를 정확하게 제공하기 위해서는 현재의 단순 키워드 중심에서 비교되는 과제정보의 범위, 비교 항목의 비중, 분석방법 등 유사성 측정 전체 프로세스 차원에서 개선방안을 살펴볼 필요가 있다.

따라서 본 논문에서는 유사·중복과제 수행의 사전방지를 위한 검색단계를 크게 3단계로 정의하였다. 1단계는 단순 키워드에 의한 유사도 검색, 2단계는 복합 키워드에 의한 유사도 검색, 3단계는 의미기반 유사도 검색이다. 이번 연구의 목표는 전체 3단계 중 실무적 측면과 현실적 측면을 고려하여 2단계에 접근하여 유사·중복과제 추출방법과 분석방법을 개선하여 Document Vector를 기반으로 한 검색엔진에 반영하여 유사도의 정확도가 1단계와 비교하여 얼마나 개선되는지 실증적 검증을 통하여 알아보고 유사도 측정 개선모형을 제안하고자 한다.

2. 관련연구

2.1. 국가연구개발 유사과제검색서비스 운영현황

정부에서는 국가연구개발 과제 기획 및 선정 시 동일하거나 유사한 연구 과제를 각 부처에서 중복 제안하여 지원 받는 사례를 사전에 방지하기 위하여 국가연구개발사업 관리 등에 관한 규정 제7조 제10항에서 연구과제 협약 전에 NTIS를 통해 과제 유사성 검토를 의무화하고 있다. 현재, NTIS에서는 2002년부터 국가연구개발사업 조사·분석을 통해 수집된 연구개발을 수행하는 전 부처의 과제정보와 기획하고 있는 과제와 비교한 유사도 정보를 제공하고 있다. 국가연구개발 과제에 대한 유사도검색 프로세스는 (그림 1)과 같다.



(그림 1) NTIS 유사과제 검색 프로세스

국가연구개발 과제를 수행하는 대부분의 부처·청 및 과제관리기관에서는 과제 선정평가 이전 단계에서 NTIS를 통해 유사과제 검색결과자료를 제출토록 하고 있으며, 과제 선정 평가단계에서 중복과제 여부를 판단하는데 참고자료로 활용되고 있다.

2.2. 유사과제 식별을 위한 알고리즘

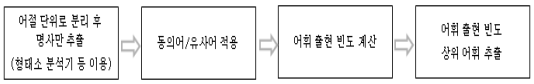
문서간의 유사성을 측정하는 기본적인 원리는 비교 대상인 두 문서에서 나타나는 어휘를 식별하여 동시에 출현하는 어휘의 빈도가 얼마나 높은가를 계산하는 것이다. 문서를 서로 비교하는 보편적인 방식으로는 현실적으로 전체 문서가 보유한 모든 어휘를 비교하지 않으며 해당 문서를 대표할 수 있는 어휘를 일부 추출하는 것이다. 대표어휘는 해당 문서에서 출현 빈도가 높은 어휘를 기준으로 추출하게 되는데, 동의어나 유사어 사전을 적용하여 같은 명칭을 서로 다르게 사용하는 경우에도 유사 문서를 찾을 수 있도록 한다.

유사과제 검색은 유사 문서 검색과 동일한 원리로 동작하지만 과제 간 비슷한 어휘가 많이 사용되는 연구과제의 특성을 고려하여 동사, 형용사 등 일반적인 어휘는 유사할 경우가 많기 때문에 제거한다.

최근의 학문적인 연구는 안정은(2010), 박동진(2009), 고방원(2010), 하정요(2008)가 형태학적 특성 기반의 유사도 검색 알고리즘을 연구하였으며, 지정훈(2009), 황인수(2009), 조정현(2009), 조혜정

(2009)은 유사도를 이용한 표절검사 알고리즘에 대해 연구하였다.[1][2][13][14][4][5][10][15] 강보영(2011)과 김윤종은 클러스터링 기법과 데이터마이닝 기법을 활용한 유사도 알고리즘에 대해 연구하였다. 이홍주(2008)가 실제 비즈니스 프로세스를 시멘틱 프로세스로 표현하고 유사한 프로세스를 검색하는 알고리즘을 연구하였다.[8] 유사과제 식별을 위한 알고리즘은 현재 이론적 배경이 진행되어 있고, 일부 시스템도 구축되어 있으나 정확도 개선을 위한 다양한 연구가 되어 있지 않은 실정이다. 앞으로 국가연구개발에 대한 투자가 지속적으로 증가되는 추세에서 투자효율성을 제고하는 것이 관건이기 때문에 유사·중복과제 수행을 사전에 방지할 수 있도록 과제 유사도에 대한 정확도를 개선하기 위한 알고리즘 연구가 필요하다.

기존 연구에서는 단일 문서 또는 과제를 대상으로 색인화하여 검색 알고리즘을 적용한 반면, 본 논문에서는 유사과제 식별에 대한 정확도를 개선하기 위해 과제와 직접적인 연관성을 가지고 있는 연구보고서 초록과 과제정보의 연구주제 키워드 정보를 분석하여 생성한 연구주제망을 검색알고리즘에 적용하였다. 본 논문에서 사용한 검색엔진 FAST는 과제의 색인어와 빈도수를 추출한 Document Vector를 기반으로 한 검색방식을 사용한다.



(그림 2) Document Vector 기반 유사도 계산 원리

FAST에서 제공하는 유사도 계산방법들을 정리하면 다음과 같다.

① Phrase를 사용하는 방법 : 항목별 가중치를 부여
(예) title:string("신종플루“, mode="phrase", weight=5)

② XRANK를 사용하는 방법 : 각 항목별 가중치를 부여
(예) XRANK(*,title:신종플루,boost=1000) and XRANK(*,title:백신,boost=1000)

③ SIMILAR_TO를 사용하는 방법 : 키워드들 확장할 수 없음. 문장검색

2.3. 과학기술 연구주제망

과학기술 연구주제망은 한국과학기술정보연구원(KISTI, Korea Institute of Science & Technology Information)이 보유한 학술정보의 주제키워드(keyword) 정보를 분석하여 주제 간의 연관 관계를 망(network)적으로 구현하는 것으로 과학기술 어휘지능망(U-WIN,User-Word Intelligent Network)과의 사상 체계 형성, 문서와의 연결 등 다양한 확장성을 가진다.(최호섭, 2007)[16]

최근의 학문적인 연구는 류창건(2008)이 한글 말뭉치를 이용한 한글 표절 탐색 모델을 연구하였고, 백종범(2009)은 키워드 불일치에 의한 정보 누락을 최소화하기 위해 대체어 후보 추출 방법을 제안하였다.[6][9] 과학기술 연구주제망은 같은 내용이라 할지라도 서로 다른 과제로 식별될 수 있기 때문에 이러한 문제점을 최소화하기 위해 사용한다. 본 논문에서는 동일한 맥락으로 NTIS가 보유한 과제정보의 주제키워드(keyword) 정보를 2개씩 쌍으로 출현하는 빈도를 계산하여 키워드간 밀접도를 구하고 연구주제간의 연관관계를 분석하여 과제 연구주제망을 구현하였다.

2.4. 가중치 부여 및 키워드 기반 유사성 검색

지금까지 유사성 검색은 단순 키워드 중심으로 연구가 진행되어 왔고, 복합 키워드 중심의 연구는 다양하지 않은 실정이다.

최근의 학문적인 연구는 안정은(2010), 박동진(2009)이 단순 키워드 기반의 유사성 검색 방안을 연구하여 연구보고서와 특허 문서(단순키워드)단위의 검색알고리즘을 제안하였다.[1][2] 고방원(2010)은 패턴 매칭 기반으로 유사도를 평가하는 시스템을 제안하였다.[13] 하정요(2008)는 색상과 형태를 이용한 내용 기반 영상 검색방법을 연구하였다.[14] 그간 다양한 분야에서 유사성 검색을 위해 단순 키워드 중심으로는 연구가 진행 되었으나 복합키워드 기반의 유사성 검색 연구는 없는 실정이다.

본 논문에서는 복합키워드를 활용한 과제의 유사도 검색을 위해 과제정보로부터 복합키워드를 추출하여 과제의 연구주제망을 생성하였고, 검색엔진에 연구보고서 초록을 추가하였다. 이를 통해 단순 키워드 기반의 유사도 검색결과와 복합 키워드 기반의 유사도 검색결과가 어떻게 달라지는지 살펴보았다.

또한, 유사도 측정을 위해 추출되는 항목에 대한 중요도가 상이하기 때문에 유사도를 계산할 때

항목별 가중치를 부여하였다.

본 논문에서는 과제정보로부터 추출되는 항목에 대한 가중치는 기존방식대로 적용하였고, 논문 초록 키워드에 대한 가중치는 과제항목에서 연구목적이나 연구내용과 동일한 비중의 가중치를 부여하였다.

<표 1> 항목별 가중치 부여 기준

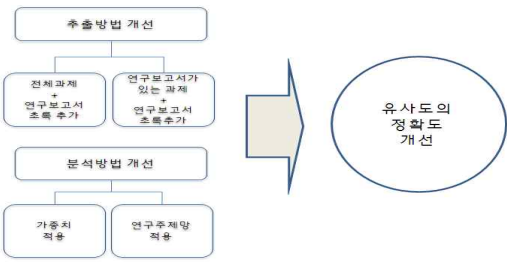
항목명	가중치	항목명	가중치
과제명	5	연구책임자	1
연구목적	3	연구내용	3
기대효과	3	키워드	1
연구보고서 초록	3		

3. 과제 유사도 측정 개선 모형 설계

3.1. 연구모형

본 논문에서는 과제의 유사도 정확도 개선에 단순 키워드가 될 것이라는 기존 연구를 바탕으로 추출방법과 분석방법을 개선하여 유사도 정확도 개선에 영향을 미치는지 살펴보았다.

과제 유사도의 정확도 개선 모형은 크게 2단계로 구분할 수 있다. 1단계는 추출방법을 과제 중심에서 연구보고서 초록까지 확대하였다. 2단계는 가중치와 과제의 연구주제망을 적용하여 분석방법을 개선하였다.



(그림 3) 유사도의 정확도 개선 모형

3.1.1 추출방법 개선

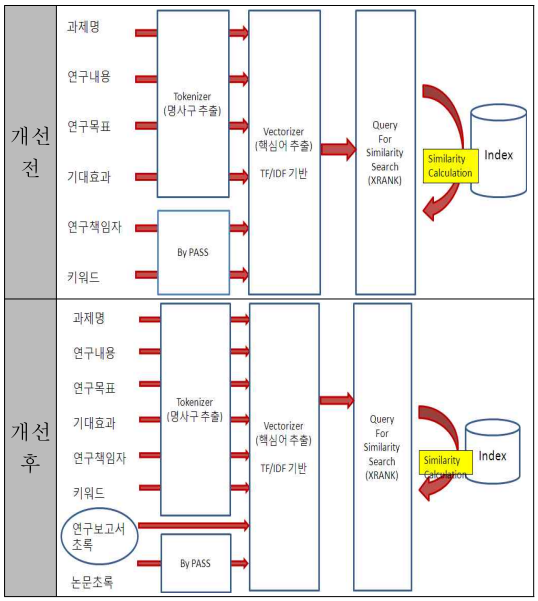
추출방법 개선은 기존에 과제정보에서만 키워드를 추출하던 것을 연구보고서 초록까지 확대하여 키워드를 추출하였다.

<표 2> 추출방법 개선 전·후 비교

개선 전	· 과제정보에서 키워드 추출
개선 후	· 연구보고서 초록까지 확대하여 키워드 추출

Document Vector(Term Vector Model)를 기반으로 한 FAST 검색엔진에 연구보고서 초록의 키워드를 추출하여 색인에 추가하였다.

<표 3> 연구보고서 초록 추가 전·후 비교



3.1.2 가중치와 과제의 연구주제망을 이용한 분석방법 개선

분석방법 개선은 NTIS 과제정보를 중심으로 연구주제 키워드(keyword) 정보를 분석하여 키워드간 가중치를 부여(weighting)하고 연구주제간의 연관 관계(밀접도)를 구현한 과제의 연구주제망을 생성하였다. 또한 추출되는 항목의 중요도가 상이함을 고려하여 항목별 가중치를 부여하였다.

본 논문에서 가중치와 과제의 연구주제망을 이용한 분석방법 개선내용은 아래 <표 4>와 같다.

<표 4> 분석방법 개선 전·후 비교

구분	가중치	연구주제망
개선전	과제명(5), 연구책임자(1),연구목적(3), 연구내용(3),기대효과(3), 키워드(1)	-
개선후	과제명(5), 연구책임자(1), 연구목적(3), 연구내용(3), 기대효과(3),키워드(1),연구보고서 초록(3)	과제의 연구주제망 활용

3.2. 가설

구분		가설
S1		추출방법을 개선한 연구보고서 초록은 유사도 정확도 개선에 영향을 미치지 않을 것이다.
	S11	과제정보 전체에 연구보고서 초록을 추가 전·후 유사도는 동일하다.
	S12	연구보고서가 있는 과제정보에 초록을 추가 전·후 유사도는 같다.
S2		분석방법을 개선한 항목별 가중치 부여는 유사도 정확도 개선에 영향을 미치지 않을 것이다.
	S22	과제정보 전체에 연구보고서 초록을 추가한 다음 항목별 가중치 적용 전·후 유사도는 동일하다.
	S23	연구보고서가 있는 과제정보에 초록을 추가한 다음 가중치 적용 전·후 유사도는 동일하다.
S3		분석방법을 개선한 가중치 부여, 연구주제망 적용 은 유사도 정확도 개선에 영향을 미치지 않을 것이다.
	S31	과제정보 전체에 연구보고서 초록을 추가한 다음 가중치와 연구주제망 적용 전·후 유사도는 동일하다.
	S32	연구보고서가 있는 과제정보에 초록을 추가한 다음 가중치와 연구주제망 적용 전·후 유사도는 동일하다.

3.3. 변수의 조작적 정의

본 논문에서는 연구보고서 초록이 과제의 유사도 정확도 개선에 유용한지, 가중치 적용 여부에 따라 유사도에 어떤 영향을 미치는지, 과제 연구주제망을 적용했을 때 정확도를 판단할 수 있는 범위가 확대되는지를 살펴보기 위해 총 6가지 유형의 실험을 실시하였다. 실험 유형은 다음과 같다.

<표 5> 실험을 위한 분류표

과제 검색대상	과제정보 전체(a)	연구보고서가 있는 과제정보만 검색(a′)
연구보고서 초록 추가 여부	추가(b)	미추가(b′)
가중치 적용 여부	적용(c)	미적용(c′)
연구주제망 적용 여부	적용(d)	미적용(d′)

<표 6> 실험 유형

가설		유형코드	조건	유형코드	조건
추출방법 개선(S1)	S11($\overline{X01}$ - $\overline{Y01}$)	x01	a+b	y01	a+b′
	S12($\overline{X02}$ = $\overline{Y02}$)	x02	a′+b	y02	a′+b′
분석방법 개선(S2, 가중치)	S21($\overline{X03}$ = $\overline{Y03}$)	x03	a+b+c	y03	a+b+c′
	S22($\overline{X04}$ = $\overline{Y04}$)	x04	a′+b+c	y04	a′+b+c′
분석방법 개선 (S3,가중치, 연구주제망)	S31($\overline{X05}$ = $\overline{Y05}$)	x07	a+b+c+d	y07	a+b+c+d′
	S32($\overline{X06}$ = $\overline{Y06}$)	x08	a′+b+c+d	y08	a′+b+c+d′

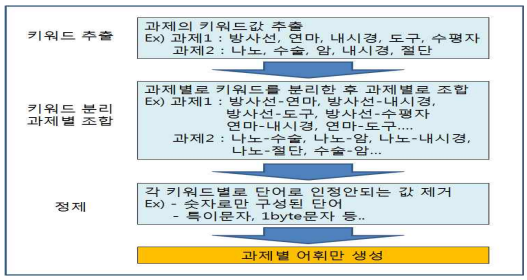
3.4. 연구의 표본

본 논문에서는 제안한 과제 유사도 측정 개선 모형을 실험하기 위해 2002년부터 2009년까지 국가연구개발 조사·분석을 통해 연구개발 관련 전 부처·청으로 수집된 251,900건의 과제정보와 16,977건의 연구보고서 초록정보를 검색엔진에 반영하여 색인화 하였다.

<표 7> 년도별 국가연구개발 과제, 연구보고서 수집건수

년도	과제건수	연구보고서건수
2009	39,598건	2,483건
2008	37,678건	1,323건
2007	33,434건	2,360건
2006	32,220건	1,609건
2005	30,610건	1,556건
2004	27,169건	1,597건
2003	26,493건	2,877건
2002년 이전	24,698건	3,172건
합계	251,900건	16,977건

또한, 251,900건의 과제정보에서 연구주제 키워드(keyword) 정보를 2개씩 쌍으로 출현하는 빈도를 계산하여 키워드간 밀접도를 구하고, 연구주제간의 연관관계를 분석한 과제의 연구주제망을 생성하였다. 과제의 연구주제망 생성절차는 다음과 같다. 첫째, 과제테이블의 Keyword값 추출한다. 둘째, 과제별로 Keyword를 분리한 후 분리된 Keyword를 과제별로 조합(1단어, 2단어)한다. 마지막으로, 다음의 데이터 정제과정을 거친다. 즉, 숫자만으로 구성된 단어 제거, 1byte문자 제거(영문자, 기호 등..), 특이문자로 이루어진 단어((1,(B,(p,-6,,1-, 2-, (m), 각 단어에 있어서는 안될 문자를 포함한 단어제거(?, <, >, ;, (,), *, #), '-' 으로 시작하거나 '-'으로 끝나는 문자의 '-' 제거, 길이가 30byte이상 되는 단어(단어가 아닌 문장일 가능성이 높음), '(단 따옴표) 치환 등을 수행한다.



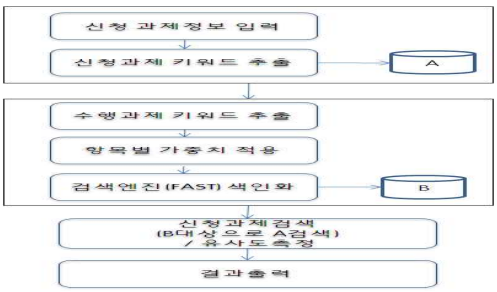
(그림 4) 과제의 연구주제망 생성 절차

그리고 분석방법을 개선하기 위해 가중치와 과제의 연구주제망을 적용하였다. 과제 유사도를 측정하기 위한 실험데이터는 2010년도 기획 및 신청과제 50건을 대상으로 하였다. 본 논문에서 사용한 서버, 데이터베이스, 검색엔진, 개발언어는 <표8>과 같다.

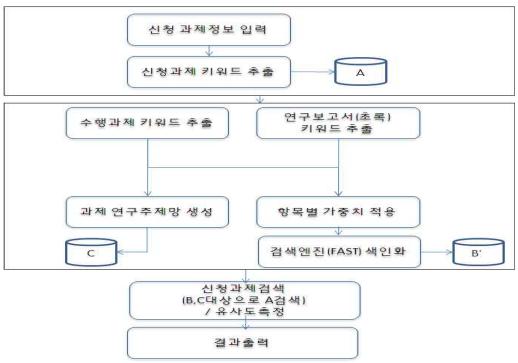
<표 8> 실험환경

구분	내용
서버	SUN Enterprise 3000
OS	Solaris
데이터베이스	Oracle 10g
검색엔진	Document Vector 기반의 FAST
개발언어	JAVA

실험은 크게 3가지 유형으로 구분하였다. 첫 번째는 아래 (그림 5)와 같이 추출방법과 분석방법을 기존 방식대로 수행과제에서 키워드를 추출(A)하고, 항목별 가중치를 적용하여 검색엔진에 색인화(B)하고 이를 대상으로 신청과제를 검색하여 유사도를 측정한 모형이다.



(그림 5) 기존 유사도 측정 모형



(그림 6) 개선 모형1

두 번째는 아래 (그림 6)과 같이 추출방법과 분석방법을 개선한 모형1이다. 과제정보와 연구보고서 초록 정보에서 키워드를 추출한 후, 항목별 가중치를 적용하여 검색엔진에 색인화(B′)하고 과제정보에서 복합키워드를 추출하여 생성한과제의 연구주제망(C)을 대상으로 신청과제를 검색하여 유사도를 측정한 모형이다.

세 번째는 두 번째 모형에서 과제와 연구보고서가 1:1로 매핑되는 정보만 추출한 모형이다. 이 실험 모형에서는 연구보고서 초록이 유사도 측정에 어느 정도 영향을 미치는지를 살펴보았다.

4. 분석방법 및 연구결과

본 논문의 실증 분석을 위한 통계분석은 SPSS 12.0을 이용하여 분석하였다.

4.1 추출방법 개선에 따른 과제의 유사도 정확도 개선

본 논문에서는 연구보고서 초록이 과제의 유사도 정확도 개선에 유용한지를 알아보기 위해 크게 2가지 유형(S11, S12)로 구분하여 실험하였다. S11은 2002년부터 2009년까지 전체 과제정보와 연구보고서 초록정보를 바탕으로 키워드를 추출하고, 색인화 작업을 통해 검색엔진에 반영하였다. S12는 2002년부터 2009년까지 과제정보와 연구보고서 정보를 비교하여 1:1로 매핑되는 경우의 정보만 추출하고, 색인화 작업을 통해 검색엔진에 반영하였다. S11 실험에 대한 대응표본 검정표를 살펴보면, 연구보고서 초록정보가 추가된 경우 과제 유사도가 평균 0.19이상 높아졌다. 또한 검정통계량 t값이 2.674이고, 이보다 작은 t값을 가질 수 있는 확률, 즉 유의수준 0.008로서 이 값은 유의수준 0.05보다 작으므로 연구보고서 초록정보가 과제 유사도를 측정하는 데 있어 영향을 미친다고 할 수 있다.

<표 9> 연구보고서 초록 추가 전·후($\overline{X01}$: $\overline{Y01}$) 비교

초록	평균	표본수	표준편차
포함후(x01)	23.6339	900	18.31601
포함전(y01)	23.4484	900	18.25612

<표 10> S11 가설 검정결과

초록	대응차					t	자유도	유의확률
	평균	표준편차	평균의 표준오차	차이의 95% 신뢰구간				
				하한	상한			
포함 전・후	.18559	2.08241	.06941	.04936	.32182	2.674	899	.008

S12 실험에 대한 대응표본 검정표를 살펴보면, 연구보고서 초록정보가 추가된 경우 과제유사도가 평균 1.82이상 높아졌다. 또한 검정통계량 t값이 10.676이고, 이보다 작은 t값을 가질 수 있는 확률, 즉 유의수준 0.000로서 이 값은 유의수준 0.05보다 작으므로 연구보고서 초록정보가 과제유사도를 측정하는 데 있어 영향을 미친다고 할 수 있다. S11과 비교했을 때 약 9배 정도 유사도가 높게 나타났고, 실험으로 통해 연구보고서 초록정보가 절대적으로 영향을 미치고 연구보고서 정보가 많이 축적될수록 유사도를 제고할 수 있다는 것을 확인할 수 있었다.

<표 11> 연구보고서가 있는 과제 대상 ($\overline{X02}$: $\overline{Y02}$)비교

초록	평균	표본수	표준편차
포함	16.1249	816	12.27259
미포함	14.3092	816	10.75306

<표 12> S12 가설 검정결과

초록	대응차					t	자유도	유의확률
	평균	표준편차	평균의 표준오차	차이의 95% 신뢰구간				
				하한	상한			
포함 전・후	1.8157	4.8584	.17008	1.4819	2.1496	10.676	815	.000

4.2 가중치와 과제의 연구주제망을 이용한 분석방법 개선 후 정확도 개선

본 논문에서는 가중치 적용 여부 및 과제 연구주제망이 과제 유사도의 정확도를 개선할 수 있는지 살펴보기 위해 크게 4가지 유형(S21, S22, S31, S32)으로 구분하여 실험하였다. 우선 가중치 반영 여부에 따라 S21, S22로 구분하였고, 가중치와 과제의 연구주제망을 동시에 반영 여부에 따라 S31, S32로 구분하였다.

S21은 2002년부터 2009년까지 전체 과제정보와 연구보고서 초록정보를 바탕으로 키워드를 추출하여 색인화 작업을 수행하였다. 그리고 항목별 가중치의 미적용군과 적용군을 분리하여 실험하였다. S22는 2002년부터 2009년까지 과제정보와 연구보고서 정보를 비교하여 1:1로 매핑되는 경우의 정보만 추출하여 색인화 작업을 수행하였다. 그리고 항목별 가중치의 미적용군과 적용군을 분리하여 실험하였다.

S31는 2002년부터 2009년까지 전체 과제정보와 연구보고서 초록정보를 바탕으로 키워드를 추출하여 색인화 작업을 수행하였다. 그리고 가중치와 과제의 연구주제망을 모두 적용하여 실험하였다. S32는 2002년부터 2009년까지 과제정보와 연구보고서 정보를 비교하여 1:1로 매핑되는 경우의 정보만 추출하여 색인화 작업을 수행하였다. 그리고 가중치와 과제의 연구주제망을 모두 적용하여 실험하였다.

S21 실험결과, 가중치를 적용하기 전보다 적용 후에 유사도가 평균 0.99이상 높아졌다. 또한 검정통계량 t값이 5.838이고, 이보다 작은 t값을 가질 수 있는 확률, 즉 유의수준 0.000로서 이 값은 유의수준 0.05보다 작으므로 가중치가 과제 유사도를 측정하는데 영향을 미치고 있다는 것을 확인할 수 있었다.

<표 13> 가중치 추가 전・후($\overline{X03}$: $\overline{Y03}$) 비교

가중치	평균	N	표준편차	평균의 표준오차
적용	24.1074	978	10.55156	.33740
미적용	23.1144	978	9.77641	.31262

<표 14> S21 가설 검정결과

가중치	대응차					t	자유도	유의확률
	평균	표준편차	평균의 표준오차	차이의 95% 신뢰구간				
				하한	상한			
적용 전・후	.9930	5.3196	.17010	.65919	1.32681	5.838	977	.000

S22 실험에 대한 대응표본 검정표를 살펴보면, 연구보고서 초록정보가 추가된 경우 과제유사도가 평균 3.43이상 높아졌다. 또한 검정통계량 t값이 22.655이고, 이보다 작은 t값을 가질 수 있는 확률, 즉 유의수준 0.000로서 이 값은 유의수준 0.05보다 작으므로 가중치가 과제유사도를 측정하는 데 있어 영향을 미친다고 할 수 있다.

S21과 비교했을 때 약 3.5배 정도 유사도가 높게 나타났고, 실험으로 통해 연구보고서 초록정보가 절대적으로 영향을 미치고 연구보고서 정보가

많이 축적될수록 유사도를 제고할 수 있다는 것을 확인할 수 있었다.

<표 15> 연구보고서가 있는 과제대상($\overline{X04}$: $\overline{Y04}$) 비교

가중치	평균	N	표준편차	평균의 표준오차
적용	17.5062	922	6.74265	.22206
미적용	14.0674	922	6.75625	.22251

<표 16> S22 가설 검정결과

가중치	대응차					t	자유도	유의확률
	평균	표준편차	평균의 표준오차	차이의 95% 신뢰구간				
				하한	상한			
적용 전・후	3.4388	4.6091	.1518	3.1409	3.7367	22.65	921	.000

S31 실험결과, 연구주제망을 적용하기 전보다 적용후에 유사도가 평균 9.25이상 낮아졌다. 또한 검정통계량 t값이 23.785이고, 이보다 작은 t값을 가질 수 있는 확률, 즉 유의수준 0.000로서 이 값은 유의수준 0.05보다 작으므로 연구주제망아 과제 유사도를 측정하는데 영향을 미치고 있다는 것을 확인할 수 있었다.

<표 17> 연구주제망 추가 전・후($\overline{X05}$: $\overline{Y05}$) 비교

의미망	평균	N	표준편차	평균의 표준오차
적용전	17.9567	875	13.24091	.44762
적용후	8.7202	875	9.42624	.31866

<표 18> S31 가설 검정결과

의미망	대응차					t	자유도	유의확률
	평균	표준편차	평균 표준오차	차이의 95% 신뢰구간				
				하한	상한			
적용 전・후	9.2365	11.4868	.3883	8.4743	9.9986	23.78	874	.000

S32 실험에 대한 대응표본 검정표를 살펴보면, 연구주제망 추가된 경우 과제유사도가 평균 8.76 이상 낮아졌다. 또한 검정통계량 t값이 23.736이고, 이보다 작은 t값을 가질 수 있는 확률, 즉 유의수준 0.000로서 이 값은 유의수준 0.05보다 작으므로 연구주제망 과제유사도를 측정하는 데 있어 영향을 미친다고 할 수 있다.

<표 19> 연구보고서가 있는 과제대상($\overline{X06}$: $\overline{Y06}$) 비교

의미망	평균	N	표준편차	평균의 표준오차
적용전	17.7634	835	11.99326	.41504
적용후	8.8766	835	9.44065	.32671

<표 20> S32 가설 검정결과

의미망	대응차					t	자유도	유의확률
	평균	표준편차	평균의 표준오차	차이의 95% 신뢰구간				
				하한	상한			
적용 전・후	8.8868	10.8064	.37397	8.1527	9.6208	23.76	834	.000

기존에 유사도는 단순 키워드 매칭으로 산출하였고, 본 논문에서 제안하는 유사도 검색방식은 복합 키워드 기반의 연구주제망을 활용함으로써 유사도에 대한 정확도를 판단할 수 있는 범위가 확대되었다.

4.3 연구결과

본 논문에서 제시한 가설들을 실증적 자료분석을 통한 검정결과를 <표 21>에 정리하였다.

<표 21> 가설의 검정

번호	가설	t 값	p 값	채택여부
S11	$\overline{X01} = \overline{Y01}$	t=2.674	p=.008	기각
S12	$\overline{X02} = \overline{Y02}$	t=10.676	p=.000	기각
S21	$\overline{X03} = \overline{Y03}$	t=5.838	p=.000	기각
S22	$\overline{X04} = \overline{Y04}$	t=22.655	p=.000	기각
S31	$\overline{X05} = \overline{Y05}$	t=23.785	p=.000	기각
S32	$\overline{X06} = \overline{Y06}$	t=23.736	p=.000	기각

5. 결론 및 향후과제

본 논문에서는 유사・중복과제 수행의 사전방지를 위한 검색단계 중 2단계에 접근하여 유사도의 정확도 개선을 위한 방안을 제안하였다. 실험결과, 단순 키워드에 의한 유사도 검색(1단계)보다 복합 키워드기반의 유사도의 정확도가 개선되는 것을 입증하였다. 또한 과제 수행 이후 산출물인 연구보고서와 과제의 연구주제망이 유사도의 정확도에 영향을 미친다는 사실을 확인하였다. 특히 본 논문에서 제안하는 유사도 검색방식은 복합 키워드 기반의

연구주제망을 활용함으로써 유사도에 대한 정확도를 판단할 수 있는 범위가 확대된 것을 확인하였다.

향후 연구해야 할 과제로는 연구보고서 원문을 추가하여 과제 유사도의 정확도를 향상시키는 것과 2개 키워드 기반의 연구주제망을 N개 키워드 기반의 연구주제망을 확대하여 과제 유사도에 대한 정확도를 판단할 수 있는 범위를 확대할 필요가 있다. 또한, 향후 연구주제망 적용 여부에 따라 유사도가 향상된 경우와 유사도가 낮게 나타난 경우에 대해서는 심층적으로 분석해 볼 필요가 있다.

그리고 추출 항목에 대한 가중치 적용이 유사도에 많은 영향을 미치고 있기 때문에 항목별 가중치에 대한 정확도 연구와 중복에 대한 판단을 객관적으로 할 수 있도록 유사도 정도에 따라 수준을 분류할 수 있는 체계를 마련할 필요가 있다. 또한 3단계에 해당하는 시소러스 기반으로 유사도 검색 프로세스를 개선하면 과제 유사도의 정확도가 제고될 것이라고 기대된다.

참 고 문 헌

[1] 안정은, 윤종민, “형태학적 특성 기반의 유사문헌 검증기법을 이용한 표준특허 사례연구”, 한국정보과학회 2010 한국컴퓨터종합학술발표논문집, 2010

[2] 박동진, 최기석, 이명선, 이상태, “유사과제 파악을 위한 검색 알고리즘의 개발에 관한 연구”, 한국콘텐츠학회논문지, Vol.9 No. 11, 2009

[3] 김윤종, “데이터마이닝 기법을 활용한 대학연구센터 지원사업의 유사성 검토방안 연구”

[4] 지정훈, 우균, 조환규, “굼벨분포 모델을 이용한 표절프로그램 자동탐색 및 추적”, 정보처리학회논문지, 제16-A권 제6호, 2009

[5] 황인수, “인터넷 검색과 형태소분석을 이용한 표절검사시스템의 개발에 관한 연구”, JOURNAL OF INFORMATION TECHNOLOGY APPLICATIONS & MANAGEMENT, 제16권 제1호, 2009

[6] 류창건, 김형준, 조환규, “한글 말뭉치를 이용한 한글 표절 탐색 모델 개발”, 정보과학회지 제14권 제2호, 2008

[7] Y. Yang and X Liu, A reexamination of text categorization methods, In SIGIR-99, 1999

[8] 이홍주, Mark Klein, “유사도 알고리즘을 활용한 시맨틱 프로세스 검색방안”, 경영정보학연구 제18권 제1호, 2008

[9] 백종범, 김성민, 이수원, “특허 정보 검색을 위한 대체어 후보 추출 방법”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제15권 제4호, 2009

[10] 조정현, 정현기, 김유섭, “웹 검색과 문서 유사도를 활용한 2 단계 신문 기사 표절 탐지 시스템”, 정보처리학회논문지 B, 제16-B권 제2호, 2009

[11] 최성필, 정창후, 전홍우, 조현양, “시맨틱 구문 트리 커널을 이용한 생명공학 분야 전문용어간 관계 식별 및 분류 연구”, 한국문헌정보학회지, 제45권 제2호, 2011

[12] 강보영, 김대원, “개선된 클러스터 유사도를 이용한 범주형 데이터의 계층적 클러스터링”, 정보과학회논문지 : 소프트웨어 및 응용, 제38권 제1호, 2011

[13] 고방원, 김영철, “패턴매칭을 이용한 유사도 비교 분석”, 한국컴퓨터정보학회논문지, 제15권, 제1호, 2010

[14] 하정요, 최미영, 최형일, “색상과 형태를 이용한 내용 기반 영상 검색”, 한국컴퓨터정보학회논문지, 제13권, 제1호, 2008

[15] 조혜정, 김지은, 손채봉, 정광수, 오승준, “통계적 분석 기반 불법 복제 비디오 영상 감식 방법”, 방송공학회논문지,제14권 제6호 , 2009

[16] 최호섭, “어휘망 구축작업에서 발견되는 한국어사전의 문제와 그 해결”,국어학회 전국학술대회, 제34회, 2007

정 옥 남

2001년 : 성균관대 정보통신대학원 (석사)
2001년 : 전자계산조직응용기술사
2010년 : 숭실대학교 IT정책경영대학원 박사과정 재학중
1989년~1997년: LG전자
1997년~현 재: 서울시, 국가과학기술위원회 사무관
관심분야 : SW공학, EA, 감리, 프로젝트관리 등

류 성 열

1980년 : 연세대학교 산업대학원 (석사)
1996년 : 아주대학교 대학원 (공학박사)
1981년~현재 : 숭실대학교 컴퓨터학부 교수
관심분야: SW 요구공학, SW 유지보수, 오픈소스 SW

김 종 배

2002년 8월 : 숭실대학교 대학원 석사
2006년 8월 : 숭실대학교 대학원 박사
2001년-현재 : (주)이엔터프라이즈 대표이사
<관심분야>소프트웨어 개발 방법론, 에이전트 시스템