국가 R&D 정보 유사문서 검색에 대한 연구

한희준, 주원균, 석중호, 최기석
한국과학기술정보연구원 NTIS센터
e-mail:hhj@kisti.re.kr

A Study on Similar Document Retrieval for National R&D Information

Hee-Jun Han, Won-Kyun Joo, Jung-Ho Seok, Kiseok Choi
NTIS Center, Korea Institute of Science and Technology Information

요 약

국가과학기술지식정보서비스(NTIS)는 국가 R&D와 관련된 과제, 성과, 인력, 시설·장비, 기술산업 정보에 대해 이용자에게 통합검색서비스를 제공한다. 이용자는 질의어를 입력하여 원하는 정보를 선별하게 되고, 한 건의 상세 메타정보 및 원문을 검색서비스의 최종 목적지로 삼는다. 이 때 이용 중인 정보와 유사한 다른 유형의 R&D 정보를 함께 제공한다면 이용자의 검색 및 탐색노력을 줄임으로써 정보획득의 요구를 쉽게 충족시킬 수 있다. 본 논문에서는 국가 R&D 정보의 메타데이터와 검색엔진의 부스팅 기법을 이용하여 이종 정보간 유사문서 검색 방법에 대해 논한다. 이는 이용자가 원하는 정보를 서비스 최종 화면(메타 상세보기)에서 제공함으로써 검색 서비스의 효율성을 증대시킨다.

1. 서론

검색서비스 이용자는 원하는 자료를 찾기 위해 질의어를 입력하고 검색시스템에 의해 제시된 결과 리스트를 탐색하면서 원문 또는 상세정보를 이용한다. 검색시스템이 질의어를 입력 받아 결과를 리턴하는데 있어서 속도 및 안정성의 문제뿐만 아니라 방대한 양의 비정형화 혹은 정형화된 문서들로부터 얼마나 정확한 결과를 제시하는가에 그 효율성이 판단된다[1]. 사용자가 원하는 결과를 상위에 랭킹시키기 위한 검색시스템의 성능 개선에 대한 요구는 사용자의 결과 문서 탐색의 노력을 줄이는데 있다고 볼 수 있다. 이런 관점에서 사용자가 결과리스트를 탐색하면서 특정 관심 문서를 이용할 때 그 문서와 유사한 다른 문서를 동시에 제공한다면 사용자는 검색결과를 재탐색하거나 다시 질의어를 생성하여 검색을 수행하는데 걸리는 시간과 노력을 낭비하지 않게 된다.

본 논문에서는 NTIS에서 서비스하고 있는 과제, 참여인력, 성과, 시설·장비, 기술산업 등 국가 R&D 정보에 대하여 이종 정보간 유사문서 검색 방법에 대해 논한다. 이 때 사용자 질의어와 특정 문서를 대표하는 문서벡터를 이용한 재검색 기법과 특정 문서의 저자명, 과학기술분류코드 등을 이용한 검색결과 부스팅(boosting) 기법을 활용한다. 2장에서는 NTIS 검색서비스에 대해 기술하고, 3장에서는 제안하는 방법을 설명하며 4장에서 결론을 맺는다.

2. 관련연구

NTIS는 연구개발의 기획에서 성과 활용에 걸쳐 연구개발의 효율성을 높이기 위한 국가과학기술지식정보서비스로써 국가 R&D를 수행하는 부처 및 청, 대표전문기관과의 정보연계를 통해 과제, 참여인력, 시설·장비, 성과 등 약 400여만건의 R&D 사업 및 성과정보를 제공한다[2]. NTIS 홈페이지는 이런 정보들에 대한 통합검색 기능을 제공하는데 사용자 질의어에 대한 검색결과 리스트를 과제, 성과, 인력별로 제공하는데 그친다. 예를 들면 사용자가 질의어를 입력하여 특정 과제에 대한 상세정보를 이용한 후 해당 과제와 유사한 다른 과제, 해당 과제로부터 유발된 성과정보 또는 해당 과제의 참여인력에 대한 정보를 동시에 이용할 방법이 없다. 이런 문제점은 사용자에게 검색 및 탐색의 부담을 주게 된다. 따라서 이종 정보간 정보특성에 적합한 유사문서 결과를 검색리스트 제시 단계에서 제공함으로써 검색서비스의 효율성을 높이고자 한다.

NTIS는 통합검색서비스를 제공하기 위해 FAST ESP 검색엔진을 사용하는데 이는 빠른 검색속도, 효율적 색인관리, 정렬의 다양성, 분산환경 지원, 검색결과 그룹핑 기능 등 다양한 부가기능을 제공한다. FAST 검색엔진의 특징 중의 하나는 각 문서마다 문서를 대표하는 키워드와 1로 정규화된 가중치 값의 쌍으로 표현되는 집합인 문서벡터(Document Vector)를 추출한다는 것이다. 문서벡터는 색인 단계에서 복합어분리, 형태소분석 등의 언어처리를 거친 키워드 리스트를 이용해 TF/IDF 기반의 가중치를 계산하여 추출되며, NTIS 를 위한 색인의 경우 10개로 제한하고 있다. 문서벡터는 과제정보의 경우 과제제목, 연구목표, 내용, 기대효과, 주제어 필드에서 추출되고 논문정보의 경우 논문제목, 요약, 주제어 필드에서 추출되며 아래

와 같은 수식으로 표현된다.

$$DV_{docid} = docid : [docvec, weight]_1, \ldots, [docvec, weight]_n$$
$$n = 1, 2, 3, \ldots, 10$$

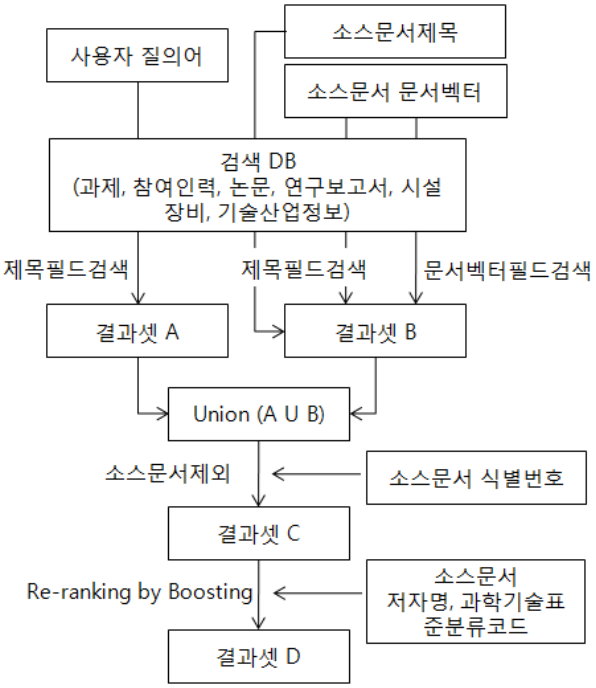
DV는 문서벡터 집합, docid는 문서의 고유식별번호, docvec는 문서벡터이고 weight는 가중치 값이다. 표 1은 NTIS 과제정보에서 추출된 문서벡터의 예이다.

<표 1> 과제정보에서 추출된 문서벡터의 예

필드명	값
과제명	소형진공유리 파일롯 양산을 위한 side sealing 및 인쇄공정기술 개발
연구목표	side sealing 및 인쇄공정 개발기술 Pilot라인 적용을 통한 진공유리 개발 , Double side sealing공법 개발 및 인쇄공법을 통한 스페이서 제조기술 개발
연구내용	Double side sealing : 로내 수소혼합가스를 사용하는 토치를 활용하여 양면 실링(실험계획법 활용 최적공정조건 도출), 스페이서 배치: 물유리와 나노 단열소재를 혼합한 페이스트의 스크린 인쇄(FEM 응력분석을 통한 배치 방법 결정)
기대효과	제조공정 핵심기술 확보 - 단열창호시스템의 핵심기술인 사이드 실링기술 및 장비기술, 지지대 제조공정 및 배치기술 확립 - 더블사이드 실링 모듈기술 : 진공이 필요한 유리 패널에 적용 가능 - 스페이서 배치 및 최적화 설계기술 : 진공창호 및 디스플레이 격벽 기술에 응용
주제어(한글)	양면, 실링, 공정, 지지대, 진공유리
주제어(영문)	double side, sealing, processing, spacer, vacuum glazing
문서벡터	**[side sealing, 1][double side, 0.866025][스페 이서, 0.866025][진공 유리, 0.866025][사이드 실링, 0.707107][인쇄 공정, 0.707107][이서 배치, 0.707107][sealing 공법, 0.5][파일롯 양산, 0.5][인쇄 공법, 0.5][유리 파일롯, 0.5][소형 진공, 0.5][vacuum glazing, 0.5][pilot 라인, 0.5][로내 수소, 0.5]**

3. 제안하는 방법

이종 정보간 유사문서 검색을 수행하기 위하여 1차적으로 사용자 질의어, 소스문서의 제목, 소스문서의 문서벡터를 조합하여 재검색이 수행되며, 재검색 결과셋은 소스문서의 저자명, 과학기술분류코드 값에 의해 부스팅 된다. 단, 정보의 종류에 따라 부스팅 기법에 사용되는 값 및 부스팅 대상의 필드가 달라질 수 있다. 여기서 부스팅은 검색결과 가운데 특정 값이 포함된 리스트를 상위에 랭킹시키는 프로세스로써 검색엔진이 제공하는 XRANK 연산이 활용된다.



(그림 1) 유사문서 검색 알고리즘

그림 1은 사용자가 질의어를 입력하여 검색을 수행한 후 검색결과 마다 유사문서를 제공하기 위한 알고리즘이다. 사용자 질의어는 사용자의 검색의도를 반영하는 중요한 요소이며 이를 이용하여 전체 타겟 DB의 제목필드에 검색을 수행하여 결과셋 A를 얻는다. 또한 소스문서의 제목과 소스문서를 대표하는 문서벡터 10개를 이용하여 타겟 DB의 제목필드와 문서벡터필드에 검색을 수행하여 결과셋 B를 얻는다. 결과셋 A와 B에서 소스문서의 식별번호를 이용하여 소스문서를 제외시킨 결과셋 C는 소스문서의 저자명, 과학기술표준분류코드에 의해 부스팅되며 재정렬된 결과셋 D는 소스문서의 유사문서 결과가 된다. 특정 값을 메타데이터로 포함한 검색결과를 상위에 랭킹시키는 부스팅 기법은 소스문서의 정보 종류 및 타겟문서의 정보 종류에 따라 다르게 정의된다. 소스문서가 과제이고 타겟문서가 참여인력 정보일 경우 부스팅 기법을 적용하기 위하여 과제의 과제책임자와 참여연구원명이 부스팅 입력값이 되며, 참여인력 정보의 성명 필드가 부스팅 대상이 된다. 즉 특정 과제의 유사문서 결과 중 참여인력 검색결과 리스트에서 그 과제의 과제책임자명이나 참여연구원명이 성명 필드에 존재하는 결과를 상위로 랭킹시킨다. 만약 소스문서가 과제이고 타겟 문서가 논문일 경우 부스팅 기법을 적용키 위해 과제의 과제고유번호가 부스팅 입력값이 되며, 논문 정보의 유발과제번호 필드가 부스팅 대상이 된다. 이 경우는 특정 과제의 유사문서 검색결과 가운데 실제 해당 과제에서 유발된 논문을 상위로 랭킹하는 부스팅 기법을 적용하는 것이다. 표 2는 소스문서 및 타겟문서 정보의 종류에 따라 적용되는 부스팅 요소에 대한 설명이다. 유사문서 검색의 활용 정보는 NTIS 를 대표하

는 과제, 참여인력, 논문, 특허, 연구보고서, 시설·장비를 대상으로 하였다.

<표 2> 정보 종류에 따른 부스팅 요소

소스문서 종류	부스팅 값	타겟문서 종류	부스팅 대상필드
과제	과제책임자명 참여연구원명	참여인력	성명
	과제책임자명 참여연구원명 과제고유번호	논문	저자명 유발과제번호
	과제책임자명 참여연구원명 과제고유번호 과학기술분류코드	특허	출원인/등록인명 유발과제번호 과학기술분류코드
	과학기술분류코드 과제고유번호	연구보고서	과학기술분류코드 유발과제번호
	과제책임자명 참여연구원명 과제고유번호	시설장비	관리책임자명 유발과제번호
참여인력	성명	과제	연구책임자명 참여연구원명
	성명	논문	저자명
	성명	특허	출원인/등록인명
	-	연구보고서	-
	성명	시설장비	관리책임자명
논문	저자명 유발과제번호	과제	과제책임자명 참여연구원명 과제고유번호
	저자명	참여인력	성명
	저자명 유발과제번호	특허	출원인/등록인명 유발과제번호
	유발과제번호	연구보고서	유발과제번호
	저자명 유발과제번호	시설장비	관리책임자명 유발과제번호
특허	출원인/등록인명 유발과제번호 과학기술분류코드	과제	과제책임자명 참여연구원명 과제고유번호 과학기술분류코드
	출원인/등록인명	참여인력	성명
	출원인/등록인명 유발과제번호	논문	저자명 유발과제번호
	유발과제번호 과학기술분류코드	연구보고서	유발과제번호 과학기술분류코드
	출원인/등록인명 유발과제번호	시설장비	관리책임자명 유발과제번호
연구보고서	과학기술분류코드 유발과제번호	과제	과학기술분류코드 과제고유번호
	-	참여인력	-
	유발과제번호	논문	유발과제번호
	과학기술분류코드 유발과제번호	특허	과학기술분류코드 유발과제번호
	유발과제번호	시설장비	유발과제번호
시설장비	관리책임자명 유발과제번호	과제	과제책임자명 참여연구원명 과제고유번호
	관리책임자명	참여인력	성명
	관리책임자명 유발과제번호	논문	저자명 유발과제번호
	관리책임자명 유발과제번호	특허	출원인/등록인명 유발과제번호
	유발과제번호	연구보고서	유발과제번호

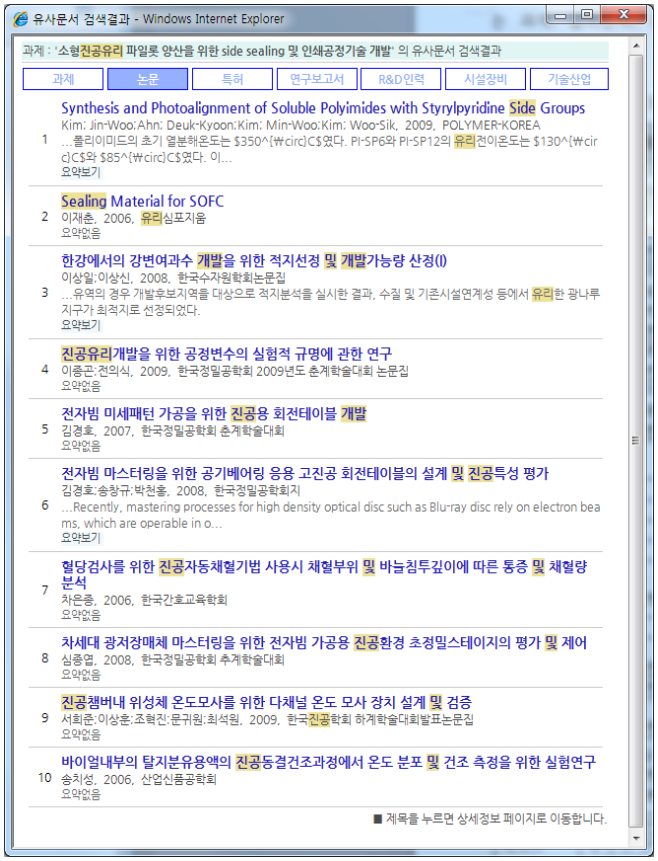
표 3은 소스문서가 과제이고 타겟문서가 특허일 경우 제안한 방법에 의해서 유사문서를 검색하기 위한 질의어 처리과정의 예이다. 사용자 질의어와 소스문서의 제목, 문서벡터 필드값을 이용해 검색을 수행한 결과리스트에 부스팅 기법을 적용하는데, 특허(검색결과)의 출원인 또는 등록인명에 과제의 책임자명, 참여연구원명이 존재하거나, 특허의 유발과제번호가 소스문서의 과제고유번호와 일치하는 결과를 검색결과 상위로 정렬한다.

<표 3> 유사문서검색 질의처리의 예

사용자질의어		진공 유리
소스문서	제목	소형진공유리 파일롯 양산을 위한 side sealing 및 인쇄공정기술 개발
	과제고유번호	1425065113
	과제책임자명	전의식
	참여연구원명	최재형, 이상호, 김을식
	과학기술분류코드	H08
	문서벡터	side sealing; double side; 스페 이서; 진공 유리; 사이드 실링; 인쇄 공정; 이서 배치; sealing 공법; 파일롯 양산; 인쇄 공법; 유리 파일롯; 소형 진공; vacuum glazing; pilot 라인; 로내 수소
질의처리단계	Q1	TITLE:string("진공 유리", mode=AND)
	Q2	TITLE:string("소형진공유리 파일롯 양산을 위한 side sealing 및 인쇄공정기술 개발", mode=OR)
	Q3	TITLE:string("side sealing double side 스페 이서 진공 유리 사이드 실링 인쇄 공정 이서 배치 sealing 공법 파일롯 양산 인쇄 공법 유리 파일롯 소형 진공 vacuum glazing pilot 라인 로내 수소", mode=OR)
	Q4	DOCVECTOR:string("side sealing double side 스페 이서 진공 유리 사이드 실링 인쇄 공정 이서 배치 sealing 공법 파일롯 양산 인쇄 공법 유리 파일롯 소형 진공 vacuum glazing pilot 라인 로내 수소", mode=OR)
	Q5	AND(Q2, Q3, Q4)
	Q6	OR(Q1, Q5)
	Q7	INVENTOR:string("전의식 최재영 이상호 김을식", mode=OR)
	Q8	RELPJT:filter("1425065113")
	Q9	SCODE:filter("H08")
	Q10	OR(Q7, Q8, Q9)
	Q11	XRANK(Q6, Q10, boostall=yes)

표 3에서 Q6 은 사용자 질의어와 소스문서의 제목, 문서벡터를 이용한 재검색을 위한 질의어이다. Q6의 수행결과 유사문서 리스트의 후보군이 결정되며, 그 이후의 질의어 처리과정은 정보의 특성에 따라 부스팅 기법을 최적화하게 된다. 표에서의 예는 과제의 과제책임자명, 연구참여자명 값을 특허의 출원인 및 등록인 필드에 검색을 수행(Q7)하고, 과제의 고유번호값을 이용해 특허의 유발과제번호 필드에 검색을 수행(Q8)하며 과제의 과학기술분류코드 값을 이용해 특허의 과학기술분류코드 필드에 검색을 수행(Q9)한 후 이것을 OR 연산 처리(Q10)한다. 최종적으

로 Q6의 수행 결과셋에서 Q10에 의해 얻어진 결과를 Q11의 연산처리를 통해 상위로 부스팅함으로써 유사문서 결과를 얻는다. 그림 2는 제안하는 방법에 의한 유사문서 검색 화면의 예시이다.



(그림 2) 국가 R&D 정보 유사문서 검색화면

4. 결론 및 향후계획

웹 이용자는 원하는 정보를 찾기 위한 첫 단계로써 검색을 수행한다. 대부분의 검색시스템은 사용자 질의어를 분석하여 사용자가 원하는 검색결과를 정확하게 제공하기 위한 나름의 프로세스를 거친다. NTIS는 국가 R&D 와 관련된 방대한 자료를 이용해 검색서비스를 제공하지만, 과제, 참여인력, 성과, 기술산업, 시설·장비 정보간 유사문서 검색 기능을 제공하여 사용자의 검색 및 탐색의 노력을 줄임과 동시에 정확하고 유용한 정보를 제공할 필요가 있다. 본 논문에서는 국가 R&D 이종 정보간 유사문서를 효과적으로 검색하는 알고리즘에 대해 제안하였다. 사용자 질의어, 소스문서의 제목, 문서벡터를 조합하고, DB 종류에 따라 저자명, 과학기술표준분류, 과제고유번호 등 부스팅 요소를 최적화하여 유사문서 이용의 효율성을 높이고자 하였다. 제안된 방법은 이용자가 원하는 정보를 찾기 위해 수많은 검색을 수행하고 탐색하는 노력을 획기적으로 줄일 것이며, 특정 정보 이용시 관련성 높은 자료를 동시에 제공함으로써 NTIS 검색서비스에 효과적으로 적용될 수 있다.

Acknowledgement

본 논문은 12년도 NTIS사업(NTIS 서비스체제 구축 및 운영)의 연구비지원에 의해 이루어졌음.

참고문헌

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

[2] www.ntis.go.kr