유서문서 및 관련연구자 검색 방법

한희준
한국과학기술정보연구원
hhj@kisti.re.kr

Similar Documents and Related Researcher Retrieval Method

Hee-Jun Han
Korea Institute of Science and Technology Information

요 약

학술정보 이용자는 연구에 필요한 자료를 획득하기 위해 검색서비스를 이용한다. 대부분의 웹 이용자는 원하는 정보를 얻기 위해 수많은 검색 질의어를 생성하여 시스템에 요청하고 선별된 정보 리스트들을 탐색하고 정보획득의 최종 목적지로써 해당 정보의 상세화면으로 이동하게 된다. 마찬가지로 논문 및 특허정보를 제공하는 학술정보서비스의 경우 이용자의 최종 목적지는 한 건의 상세 메타정보 혹은 원문이 되는데, 이 때 이용중인 정보와 유사한 다른 유형의 학술정보 및 관련 연구 분야의 연구자 제공 서비스는 이용자의 정보획득 요구를 쉽게 충족시키기 위한 필수요소이다. NDSL(국가과학기술종합정보서비스) 의 경우 동일 DB내에서의 유사문서 검색기능(논문검색에서는 유사논문 제공, 특허검색에서는 유사특허 제공)을 제공하지만 이는 이종 DB간 유사문서를 이용하고자 하는 사용자 요구사항을 만족시키지 못하는 수준이다. 본 논문에서는 논문, 특허, 연구보고서, 동향분석 자료를 포함한 학술정보 검색서비스에서 사용자 질의어와 검색엔진이 제공하는 검색 요소 및 부스팅(boosting) 기법을 이용한 이종 컨텐츠간 유사문서 리스트 및 관련 연구 분야의 연구자명 검색 서비스 기법에 대해 논한다. 이는 사용자가 원하는 학술정보를 서비스 최종 화면에서 효과적으로 제공함으로써 반복되는 검색 및 탐색의 노력을 줄일 수 있다.

1. 서 론

검색시스템은 속도 및 안정성의 문제를 넘어 비정형화 혹은 정형화된 방대한 문서들로부터 얼마나 정확한 검색결과를 사용자에게 제공해주는가에 그 효율성이 판단된다[1][2]. 웹 기반의 정보검색 서비스에서 사용자의 질의어 생성 목적은 관심있는 문서의 최종 뷰(view)에 도달하기 위한 수단이 된다. 사용자는 질의어를 생성하여 검색시스템에 결과를 요청하고, 검색시스템은 유사도 계산과 정렬 알고리즘을 포함한 저마다의 검색 프로세스를 거쳐 정제된 결과 리스트를 리턴한다. 그 후 사용자는 제시된 리스트들 중에서 검색 의도에 부합하는 문서들의 최종 상세화면으로 이동함으로써 1차 검색의 목적을 달성한다. 그러나 사용자는 이용중인 문서와 유사한 다른 문서들 혹은 관련 연구자명을 시스템이 자동으로 제시해 주길 원한다. 예를 들면 학술정보서비스에서 논문 검색결과를 이용하는 사용자는 관련된 특허나 연구보고서를 동시에 보고자 하거나 관련분야의 연구자명을 원한다. 대부분의 검색엔진은 이런 요구사항을 만족시키기 위해 문서를 대표하는 용어, 즉 문서벡터(document vector)라는 요소를 이용하여 유사문서를 추출한다[3][4]. NDSL(국가과학기술종합정보서비스)의 경우 FAST 검색엔진은 1건의 상세정보를 제공하는 화면에서 문서벡터를 이용하여 같은 종류의 DB내에서 5건의 유사문서를 제공한다. 이는 문서벡터들간의 tf, idf 값만을 이용한 유사도 계산을 거치므로, 원래 문서와의 유사성이 떨어질 가능성이 존재하며, 이종 DB간 유사문서 및 관련연구자 리스트를 이용하고자하는 요구사항을 만족시키지 못하고 있는 실정이다.

본 논문에서는 NDSL 학술정보서비스에서 논문, 특허, 연구보고서, 동향분석자료 이종 DB간 유사문서 및 관련연구자를 효과적으로 서비스하는 방법에 대해 논한다. 사용자 질의어와 검색엔진이 추출한 문서벡터를 조합하여 후보군을 생성한 후 주제분야코드와 저자명 필드값을 이용한 검색결과 부스팅(boosting) 기법을 적용해 유사문서를 제공하며, 검색엔진의 그룹핑 기능을 이용해 관련연구자를 목록화한다. 2장에서는 현재 NDSL 에서의 유사문서 서비스를 검토하고, 3장에서는 제안하는 방법을 설명하며 4장에서 결론을 맺는다.

2. 관련연구

국내 대표적인 학술정보서비스로써 KISTI에서 운영하는 NDSL을 들 수 있다[5]. NDLS은 FAST 검색엔진을 이용하여 논문, 특허, 연구보고서, 동향분석 등 약 1억여 건의 학술정보 검색서비스를 제공한다. 본 서비스의 특징 중의 하나는 상세메타정보 화면에서 유사문서를 제공한다는 것이다. 이 때 검색엔진이 문서에서 미리 추출해 놓은 문서벡터를 이용해 재검색을 수행하게 되는데, 이 문서벡터는 색인단계에서 복합어분리, 형태소분석 등의 언어처리를 거친 키워드 리스트들 중 문서에서의 출현률(frequency rate)이 높은 상위 10개의 키워드와 해당 키워드의 가중치(weight)로 구성된다. 문서벡터는 한 문서의 메타정보 가운데 제목, 요약, 주제어 필드에서 추출된

다. 한 문서를 대표하는 문서벡터는 아래와 같은 수식으로 표현된다.

$$DV_{docid} = docid : [docvec, weight]_1, ..., [docvec, weight]_n$$
$$n = 1,2,3,...,20$$

여기서 DV는 문서벡터의 집합, docid 는 문서고유번호, docvec 는 문서벡터이고 weight 는 가중치이다. 표 1은 검색엔진이 한 문서에 추출한 문서벡터의 예를 보여준다.

표 1. 문서벡터의 예

Field	Metadata
title_ko	TGIF에 의한 Human cervical cancer oncogene (HCCR) 발현 조절
title_en	TGIF Site is Involved in Expression of Human Cervical Cancer Oncogene (HCCR)
author	조광원 ; Cho, Goang-Won
publisher	Korean Society of Life Science
keywords	HCCR ; TGIF ; transcription repressor
abstract_ko	원암단백질로 알려진 Human cervical cancer oncogene (HCCR)은 발암억제 단백질인 p53과 작용하여 다양한 암조직에서 암의 유발을 촉진한다. 그러나, 아직 정확한 발암 유도기전이 알려져 있지 않다...
abstract_en	Proto-oncogene human cervical cancer oncogene (HCCR) functions as a negative regulator of p53 and contributes to tumorigenesis in various human tissues. However, it is unknown how HCCR contributes to the cellular ...
docvector	[human cervical, 1][cervical cancer, 1][cancer oncogene, 1][hccr, 0.774597][tgif, 0.591608][조절, 0.547723][mobility shift, 0.5][shift assays, 0.5][hccr promoter, 0.5][돌연 변이, 0.5]

NDSL에서는 미리 추출된 문서벡터를 이용하여 상세보기 화면에서 동일 DB에 대해 재검색을 수행함으로써 유사문서를 제공한다. 그림 1은 NDSL에서의 유사문서 검색과정이다.

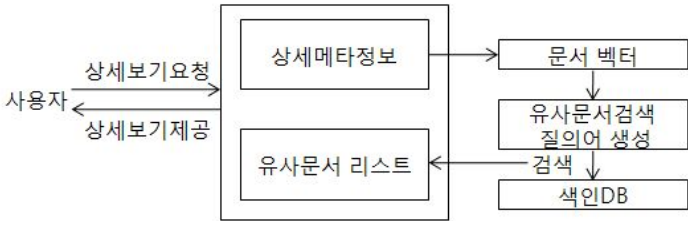

그림 1. NDSL 유사문서 검색 프로세스

사용자가 검색결과 간략 리스트에서 상세보기로 이동하면 검색시스템은 해당 문서의 문서벡터를 이용하여 유사문서 검색쿼리를 생성하고 같은 종류의 컨텐츠 색인 DB에 검색을 수행한다. 검색결과를 유사도 순으로 정렬하여 상위 5건의 문서를 유사문서 리스트로 제공한다. 이 때 유사문서를 요청하게 되는 검색쿼리는 제목, 주제어, 요약, 문서벡터 필드에 대한 OR연산이다. 아래는 유사문서 검색에 사용하는 FAST 엔진기반 질의어의 형태를 보여준다.

```
or(title:string("10 docvetors", mode=phrase,
weight=500), keyword:string("10 docvetors",
mode=phrase, weight=100), abstract:string("10
docvetors", mode=phrase, weight=100),
concepts:string("10 docvetors", mode=phrase,
weight=200))
```

이는 사용자가 작성한 질의어를 이용하지 않아 사용자의 검색의도를 반영하지 못한 검색결과를 초래하며, 논문의 주제분류코드(DDC) 또는 특허의 IPC 분류값을 이용한 검색결과 군집화 및 필터링을 통한 검색결과 정제(refine)의 개선점을 가진다.

3. 제안하는 방법

NDSL의 FAST 검색엔진은 한 문서당 10개의 문서벡터를 추출한다. 이 가운데서 0.5이상의 가중치를 가지는 문서벡터와 사용자의 원래 질의어, 사용자가 이용하는 한 건의 상세 메타정보(이하 소스문서라 함)에서 추출한 제목, 주제어를 조합하여 먼저 유사문서 후보군을 추출한다. 그 후 소스문서의 저자명 값과 주제분류코드를 이용한 부스팅 기법으로 후보군을 재정렬하는 기법으로 유사문서 검색결과를 최적화시키고자 한다. 제안하는 방법에서 사용하는 이종 DB간 유사문서 검색에 필요한 요소는 아래와 같다.

- 사용자 질의어
- 소스문서의 문서벡터 5개
- 소스문서의 문서ID, 제목, 주제어, 저자명, 주제코드

사용자가 한 건의 상세정보를 이용할 때 이종DB간 유사문서를 제공하는 알고리즘은 다음과 같다. 먼저, 사용자의 원래 질의어를 이용하여 전체 색인 DB의 제목 필드에 검색을 수행하여 결과셋 A를 얻는다. 다음으로 소스문서의 제목 값으로 전체 색인 DB의 제목에 검색을 수행한 결과와 소스문서의 주제어, 문서벡터를 조합하여 전체 색인 DB의 제목과 문서벡터 필드에 검색한 결과의 교집합 B를 구한다. 여기서 소스문서의 제목, 주제어, 문서벡터의 값의 공기(co-occurrence)는 OR 연산자로 처리한다. 그 다음 결과셋 A와 B의 합집합(Unoin set)으로부터 소스문서를 제외시키기 위해 소스문서의 문서ID를 ANDNOT 연산자 처리하여 최종적으로 유사문서 리스트에 해당하는 결과셋 C를 얻게 된다. 다음으로 결과

셋 C의 문서들 중에서 소스문서와 같은 주제분야에 해당하거나 소스문서의 저자명이 포함된 문서를 클러스터링하기 위하여 부스팅(boosting) 기법이 적용된다. 즉, 결과셋 C의 문서들을 부스팅 요소인 소스문서의 주제분류코드와 저자명 값을 이용해 재정렬하는데, 이때 FAST 엔진이 제공하는 XRANK 연산자가 적용된다. 최종적으로 얻은 유사문서 리스트에 해당하는 메타정보 중 저자명 필드 값을 그룹핑한 후 출현률이 높은 순으로 나열하여 관련연구자 리스트를 제공한다.

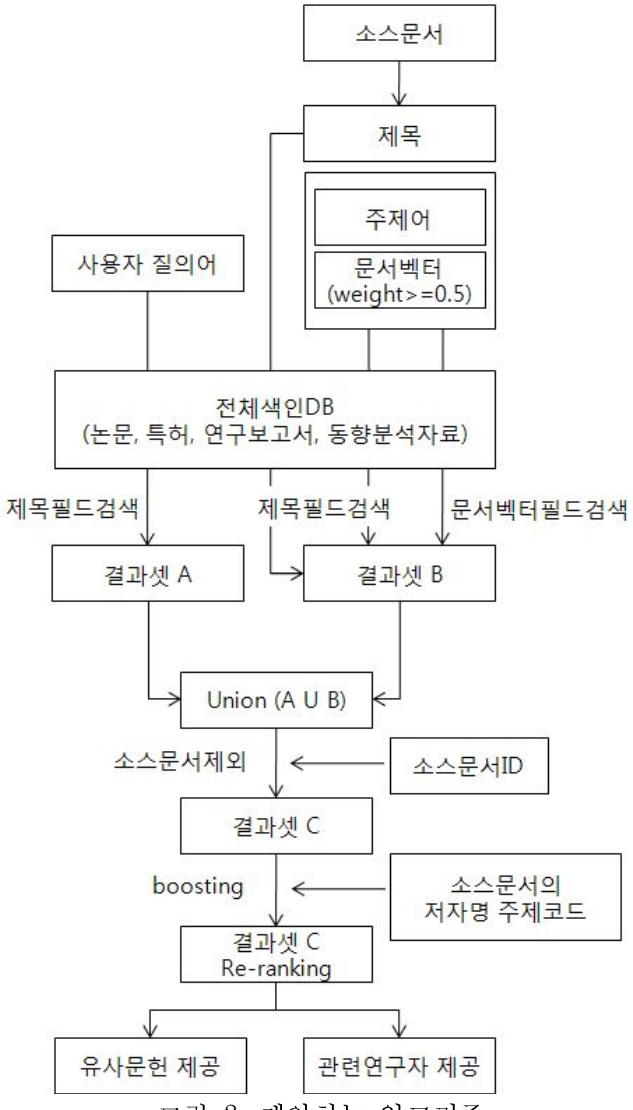

그림 2. 제안하는 알고리즘

결과셋 A는 전체 DB의 제목 필드에서 사용자 질의어를 통해 검색한 결과로서 초기의 사용자 검색의도를 반영한 결과이다. 결과셋 B는 사용자가 현재 이용하고 있는 한 건의 상세정보로부터 제목과 주제어, 문서벡터를 추출한 후 이를 전체 DB의 제목, 문서벡터필드에 검색을 수행한 결과로써, 이는 소스문서와 유사한 결과이다. 이를 머지(murge)하여 소스문서 1건을 제외한 후 후보군 결과셋 C로부터 소스문서와 유사도가 높은 리스트만을 서비스 대상에 포함시키기 위하여 부스팅 기법을 적용한다. 즉, 결과셋 C의 문서들 중에서도 소스문서의 저자명이 포함된 문서가 존재한다던가, 혹은 소스문서와 주제분야코드 혹은 특허의 경우 IPC분류가 일치하는 문서가 존재한다면 이를 상위로 부스팅한다. 동시에 결과셋 C의 저자필드로부터 저자명 그룹핑 정보를 생성하여 출현률이 높은 순으로 나열하면 제시하는 유사문서로부터 관련 저자, 관련 연구자 또는 관련 특허 출원인을 제공할 수 있다. 유사문서 검색 과정에서 질의어 처리 단계는 표 2에서 보여준다. 표 2는 한 건(소스문서)의 메타정보와 이를 이용한 유사문서 검색 질의 처리과정의 예시이다.

표 2. 유사문서검색 질의처리

사용자질의어			위암 유전자
소스문서	문서ID		JAKO200503018280686
	제목		위암조직에서의 MAGE 유전자 발현
	주제어		위암 ; MAGE ; Gastic cancer ; RT-PCR
	저자명		최재형 ; 이상호 ; Choi, Jae-Young ; Lee, Sang-Ho
	주제분류코드 (DDC)		616994
	문서벡터 (w>=0.5)		cancer tissues ; gastric cancer ; mage gene ; 위암조직 ; gastric carcinomas
질의처리단계	Q1		TITLE:string("위암 유전자", mode=AND, weight=200)
	Q2		TITLE:string("위암조직에서의 MAGE 유전자 발현", mode=OR, weight=200)
	Q3		TITLE:string("위암 MAGE Gastic cancer RT-PCR cancer tissues gastric cancer mage gene 위암조직 gastric carcinomas", mode=OR)
	Q4		DOCVECTOR:string("위암 MAGE Gastic cancer RT-PCR cancer tissues gastric cancer mage gene 위암조직 gastric carcinomas", mode=OR)
	Q5		AND(Q2, Q3, Q4)
	Q6		OR(Q1, Q5)
	Q7		ANDNOT(Q6, DOCID:filter("JAKO200503018280686"))
	Q8		OR(AUTHOR:string("최재형 이상호 Choi, Jae-Young Lee, Sang-Ho", mode=OR), DDC:filter("616906"))
	Q9		XRANK(Q7, Q8, boostall=yes)

제안하는 방법에서 사용되는 검색필드는 논문, 특허, 연구보고서, 동향분석자료에 공통적으로 존재하므로, 이종 DB간 유사문서 검색에 효과적으로 적용된다. 사용자는 특허 정보를 이용하는 동시에 관련 유사논문, 연구보고서 및 관련분야의 연구자 및 특허 출원인을 제공 받을 수 있고 인명을 통한 다른 문서로의 탐색서비스도 가능하다. 그림 3은 학술정보서비스에서 제안하는 방법에 의

한 유사문서 및 관련연구자 클러스터링 서비스 화면의 예시이다. 이는 표 2에서의 논문의 상세정보와 함께 제공되는 관련논문, 관련특허, 관련연구보고서, 관련동향분석 자료이며 동시에 관련저자, 관련발명자 등 연구자 리스트를 제공한다.

그림 3. 유사문서 및 관련연구자 서비스 화면

4. 결 론

몇몇 학술정보서비스가 사용자가 이용하는 문서에 대한 유사문서를 제공하고는 있지만, NDSL의 경우 그 범위가 같은 DB내로 국한되어 있고 유사문서의 정확성이 떨어지므로 논문, 특허, 연구보고서, 동향분석자료 등 이종 DB간 유사문서 및 관련연구자를 이용하고자 하는 요구사항을 만족시키지 못하고 있다.

본 논문에서는 이종 DB간 유사문서를 효과적으로 검색하는 알고리즘에 대해 제안하였다. 사용자 질의어, 소스문서의 제목, 주제어, 문서벡터를 조합하고, 저자명과 주제분야코드를 이용해 부스팅하는 기법을 적용하여 유사도가 높은 문서들을 서비스하도록 하였으며, 이 결과를 활용하여 관련연구자 리스트를 제공하였다. 이는 먼저 NDSL 학술정보서비스의 지능화 목적으로 NDSL-STAR 사이트에 적용되었다[6]. 제안된 알고리즘은 콘텐츠의 종류에 상관없이 다른 검색서비스에 효과적으로 적용될 수 있다.

5. 참고문헌

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

[2] Can, F., and E. A. Ozkarahan, “Dynamic Cluster Maintenance”, Information Processing & Management, Vol. 25, pp.275-291, 1989

[3] 김광영, 곽승진, “인접한 단어와 키워드 주제어 정보에 기반한 유사 문헌 검색 시스템 개발”, 한국도서관정보학회지, Vol.40, No.3, pp.367-387, 2009

[4] Fox, T.W., “Document vector compression and its application in document clustering”, Electrical and Computer Engineering, Canadian Conference on, IEEE, pp. 2029-2032, 2005

[5] http://www.ndsl.kr/

[6] http://star.kisti.re.kr/ndslstar/