벡터스페이스모델과 시소러스를 이용한 특허검색시스템의 성능향상

임성신[O] 정홍석 한기덕 권혁철
부산대학교 컴퓨터공학과
{sslim[O], hsjung, templer, hckwon}@pusan.ac.kr

Improving Patent Information Service System using Vector Space Model and Thesaurus

Sungshin Lim[O] Hongseok Jung, Gideok Han, Hyuk-Chul Kwon
Dept. of Computer Science & Engineering, Pusan National University

요 약

지적재산권이 산업의 핵심으로 자리잡음으로써 특허의 중요성이 날로 증가하고 있다. 현재 특허문서 검색을 서비스하고 있는 상용시스템의 경우 문서간의 유사도나, 질의어에 따른 순위(Ranking)가 매겨지지 않는 불리언 모델이 검색에 사용되고 있다. 본 논문에서는 유사도에 기반한 순위화가 가능한 벡터모델기반의 특허검색 시스템을 개발하고 시계분야의 시소러스를 구축하여 시계분야의 특허검색 시스템에 적용하였다. 쿼리확장의 성능을 평가하기 위해 10개의 쿼리로 실험하였고 평균 36.2%의 정확도가 향상되었다. 그리고 검색결과의 오른쪽에 시소러스를 제시함으로써 특허검색시스템을 이용하는 사용자에게 추가 질의어를 쉽게 선택할 수 있도록 하여 인터페이스 부분의 향상을 추구하였다.

1. 서 론

특허정보는 기업의 생존전략과 직결되는 정보로서 기업의 기술개발단계에서 사업화까지의 경영전략에 연계되는 중요한 정보로 사용되고 있다. UR/TRIPS타결 및 WTO 출범으로 무한기술경쟁시대에 돌입하게 됨으로써 세계 각국은 자국의 산업재산권 보호를 강화하고 있으며, 선진 기업들도 특허전략을 공격적으로 전환함에 따라 대내외적으로 특허출원 및 분쟁이 급증하고 있어 각 기업은 충분한 사전조사로 중복연구 및 특허분쟁을 예방하고, 기술개발의 동향파악 및 아이디어의 입수로 적극적인 기술개발을 통한 대응이 필요하다. 그리고 특허의 출원 건수는 매년 빠르게 증가하고 있으며 현재 전세계적으로 연간 500여만건 정도의 특허정보가 발생되고 있으며 국내에서는 연간 25만건 정도가 발생하고 있다[1][2].

현재 특허검색을 서비스하고 있는 상용시스템[3]의 경우 문서간의 유사도나, 질의어에 따른 순위(Ranking)가 매겨지지 않는 불리언 모델이 검색에 사용되고 있다.
본 논문에서는 벡터 스페이스 모델에 기반하여 순위화가 가능한 특허검색시스템을 개발하였다. 그리고 시계분야의 시소러스를 구축하여 시계분야의 특허검색 시스템에 시범적으로 적용하였다. 쿼리확장의 성능을 평가하기 위해 10개의 쿼리로 실험하였고 평균 36.2%의 정확도가 향상되었다. 그리고 검색결과의 오른쪽에 시소러스를 제시함으로써 특허검색시스템을 이용하는 사용자에게 추가 질의어를 쉽게 선택할 수 있도록 하여 인터페이스 부분의 향상을 추구하였다.

2. 시스템 구성

그림1은 특허정보서비스시스템(PISS: Patent Information Service System)의 전체구조를 나타내고 있다.

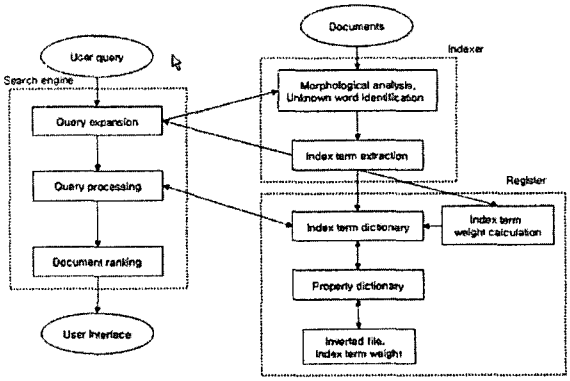

그림 1. PISS의 전체구조

PISS는 크게 3가지 모듈(검색기, 색인기, 등록기)로 구성되어 있다.

검색기는 사용자 인터페이스에서 자연언어 질의문 혹은 불리언 질의문을 입력받아 색인기를 이용하여 질의어를 생성해 내는데, 색인기는 질의어로 단일명사 혹은 복합명사를 추출해 낸다.

벡터스페이스모델에서 검색결과를 얻기 위해서는 질의문과 모든 문서와의 유사도를 구하고 유사도 값이 일정치 이상되면 검색한다. 그러나 대용량의 문서를 검색하는 시스템에서 질의문과 모든 문서의 유사도를 계산하기 위해서는 상당한 양의 처리 시간이 필요하다. 그러므로 처리 시간을 줄이기 위해 먼저 색인기에 의해 생성된 질의어들의 불리언검색(OR 연산)을 수행하여 순위를 매기지 않고 문서를 검색한 다음, 코사인 유사 계수를 이용하여 검색된 문서들에서 정보 획득이 쉽도록 하기 위해 이 문서들과 질의문의 유사도를 계산하여 순위를 매겨 출력한다.

색인기는 등록될 문서에서 색인어를 추출하고, 입력되는 질의문에서 질의어를 추출한다. 색인기는 형태소분석, 중의성 제거, 미등록어 추정, n-gram 처리 등을 이용하여 문서(혹은 질의문)에서 명사를 추출한다.

등록기는 문서를 색인기에서 넘겨받은 색인 결과를 이용하여 문서 데이터베이스로 구조화시키는 시스템이다. 문서 데이터베이스는 색인어 파일과 각 색인어에 따른 문서 번호 역파일로 구성되고, 색인기에서 넘겨받은 단어의 출현 빈도를 색인어 가중치로 삽입한다.

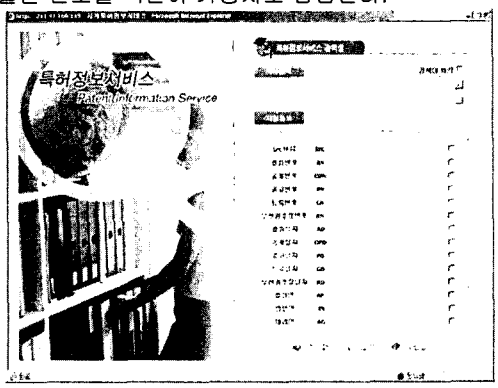

그림 2. PISS interface

2.1 시소러스 구축

시계분야의 시소러스를 구축하기 위해 국제표준의 분류체계(HS Code, SITC, NAICS, UNCCS, UNSPSC)를 분석하였다.

HS Code(harmonized Commodity Description and Coding System)는 WCO(World Customs Organization)가 수출입 상품에 대한 무역통계 및 관세행정을 위하여 개발한 코드로써 국내에서는 관세청에서 사용하고 있다. 현재 176개 국가(WCO 국제 협의회 회원국 98개, 비회원국 78개)에서 HS Code를 사용함으로써 세계 무역의 98%이상을 담당하고 있다. 하지만 관세를 위해 분류된 코드로써 논리적이지 못하고 분류기준이 애매모호하여 일반인들이 사용하기가 어렵다. 그리고 관세 이외의 변수는 모두 기타항목으로 분류되어 일반적인 제품 분류로는 적합하지 않다.

UNSPSC(UN Standard Products and Services Classification)는 D&B사와 UNDP가 함께 개발한 품목코드 표준으로 UN의 UNCCS와 Dun & Bradstreet사의 SPSC(Standard Product and Services Codes)를 결합한 것이다. UNSPSC는 세그먼트(Segment), 패밀리(Family), 클래스(Class), 코모디티(commodity)등 4단계의 계층구조를 이루며, 각 단계가 두 자리의 수로 구성되어 총 8자리 수의 코드체계이다. 글로벌 marketplace에 사용되는 품목과 서비스를 분류하는 첫번째 코드로서 전자상거래에 활용하기 위하여 8000개 이상의 품목과 서비스를 분류하고 있다. UNSPSC는 시계산업과 기본적인 체계가 맞고 전세계적으로 전자상거래의 표준으로 검토되고 있는 상황이다. 또한 수정 및 보완이 민주적이고 일정이 짧아서 의견 반영이 비교적 용이하므로 이를 기준으로 시계분야의 시소러스를 구축하였다.

시계분야의 시소러스 구축을 할 때 영어와 한국어를 쌍으로 구축하고 하나의 개념단위는 동의어 집합인 synset(set of synonym)으로 구축하였다. 그리고 특허문서는 변리사나 그 분야의 전문가가 작성을 하지만 용어의 표준화를 이루고 있지 않으므로 오용어가 많이 발생한다. 그래서 PISS의 성능 향상을 위해 시소러스를 구축할 때 오용어도 추가로 구축하였다[4][5][6][7]. 전체 구축한 규모는 311개의 synset을 1694개 단어로 구축을 하였다.

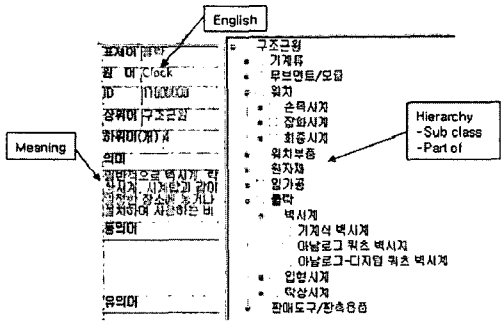

그림3. 구축한 시소러스

2.2 검색 모델

본 논문에서는 벡터 스페이스 모델(vector space model)을 이용하여 문서와 질의문을 표현한다. 벡터 스페이스 모델에서의 문서와 질의문은 단어들의 벡터로 표현될 수 있다[4]. 문서 집단이 n개의 색인어로 구성되어 있다면 문서 D_i는 다음과 같이 n차원의 벡터로 표현된다.

$$Di = <td_{i1}, td_{i2}, \cdots, td_{in}>$$

여기서 td_{ij}는 문서 D_i에서 색인어 써의 가중치 값이다. 질의어 벡터는 다음과 같다.

$$Qi = <tq_1, tq_2, \ldots, tq_n>$$

벡터 스페이스 모델은 불리언 탐색 기법의 단점을 보완하는 검색 기법으로, 불리언 검색 기법은 다음과 같은 몇가지 문제점을 가지고 있다.

첫째, 질의문과 완전히 일치되는 문서만이 검색되므로 부분적으로 일치하는 문서는 검색할 수 없다.

즉, 이용자가 자신의 정보 요구를 정확하게 표현할 수 있어야 한다는 것을 의미하므로 해당 영역에서 사용되는 검색단어에 익숙한 사용자나 그 영역에서 검색 경험이 있는 이용자에게는 적합할 수 있으나, 이용자가 항상 자신의 관심 분야의 단어에 정통한 것은 아니므로 명확하지 않은 정보 요구를 가진 이용자는 효과적인 검색을 할 수 없음을 의미한다. 둘째, 검색되는 문서를 질문과의 유사도 크기의 내림차순으로 출력할 수 없다. 셋째, 검색어로 표현되는 각 개념들을 자신의 정보 요구 형태에 따라 상대적인 중요도를 나타내지 못한다.

벡터 스페이스 모델의 장점은 다음과 같다. 첫째, 불리언 연산자를 사용할 필요가 없다. 즉, 이용자 요구를 나타내는 자연언어 문장으로부터 자동적으로 질의어를 추출하여 질의어 벡터를 형성할 수 있다. 둘째, 검색어를 자신의 정보 요구 형태에 따라 상대적인 중요도를 쉽게 나타낼 수 있다. 셋째, 유사도 함수를 이용하여 검색 결과에 순위를 부여할 수 있고, 부분적으로 일치하는 문서도 검색할 수 있다.

문서 벡터와 질의문 벡터간의 유사도(similarity)를 계산하는 여러 가지 함수 중에서 본 논문에서는 코사인 상관관계에 기반한 벡터매칭(vector matching)을 이용하여 유사도를 구한다. 코사인 유사도 함수는 일반적으로 좋은 방법으로 평가 받고 있다. 그러나 비교되는 벡터들의 곱으로 나누기 때문에 보통 길이가 긴 문서의 유사도가 낮다[8]. 코사인 유사도 함수는 다음과 같다.

$$\mathrm{Similarity}(D_j, Q_k) = \frac{\sum_{i=1}^{n} (td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^{\,2} \times \sum_{i=1}^{n} tq_{ik}^{\,2}}}$$

벡터 스페이스 모델에서 단어에 가중치를 부여함으로써 검색결과의 순위에 영향을 미친다. 여러 가지 단어 가중치 부여 방법 중에서 단어가 출현하는 빈도를 단어의 중요도와 연관시키는 단어 빈도 가중치가 일반적으로 이용된다. 즉, 출현 빈도를 단어 가중치로 이용하는 시스템에서 가중치가 높은 질의어와 일치하는 색인어를 가지고 있는 문서일수록 유사도 값이 커진다.

3. 시스템 평가

시스템은 10개의 질의어로 검색결과가 질의어와 유사한지(relevant), 유사하지 않은지(non-relevant) 평가 하였다. 검색 결과가 10개 이상일 때는 상위10개의 문서로 평가를 수행하였다. 상위 10개의 결과를 평가했을 때 Boolean model의 정확도는 평균 37.8%정도였으나 cosine measure로 유사도를 계산하여 유사도 순으로 정렬한 결과는 평균 74%의 정확도를 보이고 있다. 그리고 query를 확장함으로써 관련 문서도 증가함을 알 수 있었다.

그리고 검색결과의 오른쪽에 시소러스를 제시함으로써 특허검색시스템을 이용하는 사용자에게 추가 질의어를 쉽게 선택할 수 있도록 하여 인터페이스 부분의 향상을 추구하였다.

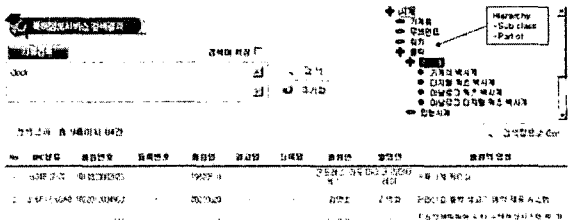

그림 4. 사용자 인터페이스의 개선

4. 결론

산업계의 기술경쟁력 강화와 지적 재산권 확보의 중요한 수단으로써 특허의 중요성은 갈수록 커져가고 있다. 현재 특허문서 검색을 서비스하고 있는 상용시스템의 경우 문서간의 유사도나, 질의어에 따른 순위(Ranking)가 매겨지지 않는 불리언 모델이 검색에 사용되고 있다. 본 논문에서는 유사도에 기반하여 순위화가 가능한 벡터모델기반의 특허검색시스템을 개발하고 시계분야의 시소러스를 구축하여 시계분야의 특허검색 시스템에 적용하였다. 쿼리확장의 성능을 평가하기 위해 10개의 쿼리로 실험하였고 평균 36.2%의 정확도가 향상되었다. 그리고 검색결과의 오른쪽에 시소러스를 제시함으로써 특허검색시스템을 이용하는 사용자에게 추가 질의어를 쉽게 선택할 수 있도록 하여 인터페이스 부분의 향상을 추구하였다.

좀더 발전된 검색 시스템이 되기 위해서는 현재 구축한 시소러스를 보완하여 웹 시소러스를 만들고 이를 기반으로 특허문서를 annotation을 수행하여 metadata를 구축한다면 시맨틱웹 기반의 서비스도 가능할 것으로 보인다.

5. 참고문헌

[1]김낙현, 정수용, 강창수, 이재황, " 웹을 이용한 특허정보 검색서비스" 한국정보처리학회논문집 제6권 제3호, 1999, 80-85.

[2]원상훈, 노태길, 손기준, 박정희, 이상조, " 특허정보 검색을 위한 벡터스페이스 검색모델의 적용" 한국정보과학회 학술발표회논문집, 2003. 10.

[3]KIPRIS(Korean Industrial Property Rights Information Service), http://www.kipris.or.kr

[4]Vinay Kakade, Madhura Sharangpani, " Imporving the Precision of Web Search for Medical Domain using Automatic Query Expansion"

[5]D. Cai, C.J.van Rijsbergen, " Automatic Query Expansion Based on Directed Divergence"

[6]M.Makita, S.Higuchi, A.Fujii, T.Ishikawa, " A System for Japa-nese/English/Korean Multilingual Patent Retrieval"

[7]W.B.Frakes and R.Baeza-Yates, " Information Retrieval: data structure and algorithms" , Printice Hall, New Jersey, 1992.

[8]G.Salton and M.J.McGill, " Introduction to Modern Information Re-trieval" , McGrawHill, 1983.