# 특허정보 검색을 위한
# 벡터스페이스 검색모델의 적용

원상훈[0] 노태길 손기준 박정희 이상조

경북대학교 컴퓨터공학과

{shwon[0], nayas, kjson, jhpark}@sejong.knu.ac.kr sjlee@knu.ac.kr

# Vector Space Model for Patent Information Retrieval System

Sang-Hoon Won[0] Tae-Gil Noh Ki-Jun Son Jung-Hee Park Sang-Jo Lee

Dept. of Computer Engineering Kyungpook National University

**요 약**

본 논문은 특허 문서에 맞게 벡터스페이스 모델을 적용하여 특허정보 검색기를 구현한다. 기존의 상용 특허 검색 시스템의 문제점을 제시하고, 특허 문헌의 특징을 분석하여, 이를 반영한 특허 문헌 검색용의 벡터 스페이스 모델을 제시한다. 하나의 특허 문서는 서로 상이한 특성을 지닌 텍스트와 데이터의 조합으로 이루어져 있다. 따라서 이를 하나의 벡터로 표현하는 것이 용이하지 않다. 이에 대해 본 연구에서는 내용 필드들을 특성에 따라 둘 이상의 벡터로 표현하고, 수치 및 고유명 필드는 불린검색형태로 처리되는 혼합형 벡터 모델을 제안한다. 각 필드의 특징에 맞게 색인어를 추출하며, 텍스트 필드의 색인어를 벡터로 표현하는 과정에서는 잘 알려진 TF-IDF 가중치를 사용하되, 특허 문서가 IPC 특허 분류 기준에 따라 완전 분류되어 있는 문서라는 특징을 이용, 보다 정확한 가중치를 부여한다. 실험과 성능평가를 통하여 제안한 특허 모델의 유용성을 보인다.

## 1. 서 론

특허의 출원 건수는 매년 증가하고 있어 현재 전세계적으로 년간 500여만건 정도의 특허정보가 발생되고 국내에서는 년간 25만건 정도가 발생하고 있다[1][2].특허 정보검색은 각 국가에서 출원된 특허정보 모두가 빠짐없이 수록된 방대한 양의 정보원에서 탐색해야 하기 때문에 이용자는 부적합한 특허정보의 검색을 억제하고 적합정보만을 선별하여 검색해야 하는 부담이 따른다[3].

현재 특허문서 검색을 서비스하고 있는 상용시스템의 경우 대부분 불린 검색에 기반을 두고 있다.문서간의 유사도나, 질의에 따른 순위(Ranking)가 매겨지지 않는 불린 모델이 아직 일반적인 특허 검색 모델인 이유는, 일반 문서를 대상으로 만들어진 기존의 벡터모델이나 확률모델 시스템을 특허 문헌의 데이터베이스에 그대로 적용하기가 쉽지 않다.

이러한 어려움은, 하나의 특허 문헌이 하나의 벡터나 수치 표현으로 옮기기 어려운 상이한 특징의 텍스트/필드들의 조합이라는 것에서 기인하다. 본 논문에서는 이러한 점들을 감안하여 벡터 모델에 기반한 특허정보검색시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 특허문서의 특징을 살피고, 3장에서는 이를 고려한 특허정보 검색시스템을 제안한다. 4장에서는 실험 및 평가를 행하고, 5장에서 결론을 맺는다.

## 2. 특허문서를 위한 벡터모델

### 2.1 특허문서의 특징과 적용의 문제점

특허문서는 기사문이나 일반 텍스트와는 상당히 다른성격을 지니고 있다. 보통의 설명문과 비슷한 텍스트로 이루어진 필드들이 있는가 하면, 특별한 양식을 갖추어 적히는 텍스트 필드도 있고, 숫자만이 의미가 있는 필드도 있다. 또,고유명사가 적히는 필드, 그리고 분야에 따라서는 도면만이 중요한 역할을 하는 문서도 있으며, 심지어 특허의 출원자가 부여한 인덱스 필드까지 모여, 하나의 단일한 특허 문서를 이룬다. 사실상 서로 다른 특징을 지닌 텍스트들이 모여 하나의 문서를 이루고 있다는 것이, 이러한 하나의 문서를 검색해내는 작업을 쉽지 않게 만든다.

일반적인 특허문서의 필드들을 살펴보면 표1과 같다. 크게 나누면 단일한 단어들로 이루어진 필드와, 텍스트가 포함된 필드가 있다. 단일한 단어/숫자로 이루어진 필드에는, 숫자와 기호로만 이루어진 필드(공개번호, 공개일, 출원번호, 출원일, IPC, 등), 한 두개의 고유명사로 이루어진 필드(출원인, 대리인, 발명인)로 나누어 볼 수 있다. 일반적인 텍스트와 가까운 필드에는, 요약필드 및 청구 범위, 그리고 특허의 내용이 기술되는 긴 내용 필드가 있다. 또 특허출원인이 검색에 도움이 되도록 자유롭게 선택하여 부여 할 수 있는 키워드 필드도 있다.

표1. 각 필드별 특징

| | |
|---|---|
| 숫자로 이루어진 필드 | 공개번호,공개일,출원번호,출원일, IPC분류 등 |
| 고유명사 필드 | 출원인,발명자 대리인 |
| 일반 텍스트형식의 필드 | 요약,키워드 |
| 특정스타일로 쓰여진 필드 | 청구범위 |

기존의 벡터 모델 검색시스템을 특허 문서에 바로 적용하기 어려운 것은, 특허 문서가 이와 같이 다양한 특성을 지는 서로 다른 텍스트의 조합이기 때문이다.

벡터모델은 하나의 문서를 이루는 단어들을 색인하고, 이러한 단어에 가중치를 부여한 다음, 문서를 하나의 벡터에 표현하여 검색을 행하나, 특허 문헌은 상이한 성격의 텍스트가 조합되어 하나의 문서를 이루기 때문에, 여기에 담긴 개별 정보의 특징을 잃지 않으면서 하나의 벡터로 환원하는 것이 쉽지 않다. 이같은 점이 특허 문헌을 위한 검색 시스템을 구현하는 데에 어려운점 이라면, 특허 문서에는 일반 문서에서는 없는, 검색에 유용한 추가정보도 포함되어 있다. 먼저 모든 특허 문서는 해당 특허가 어느 분야에 속하는지를 명시하는 IPC 분류표에 의해 완전히 분류되어 있는 문서이다. 특허 문서는 필드별로 나타날 값들을 미리 예측할 수도 있어, 색인어 추출에 유용한 정보가 되기도 한다. 예를 들면 출원인, 발명인 같은 필드에는 언제나 고유명사가 나타나며, 숫자가 포함되는 필드에는 숫자와 기호만이 등장한다. 또한 텍스트로 이루어진 필드지만, 청구범위와 같은 필드는 오랫동안 특허관련분야에서 정립된 주어진 형태의 순서로 제시되는 문장들이어서, 보통의 텍스트와는 문장의 성격과 나타나는 단어들의 중요성이 다르다.

### 2.2 특허문서 검색을 위한 벡터모델

특허 문서가 상이한 성격의 텍스트의 조합이라는 점에 대해서, 이를 하나 이상의 벡터로 표현하여 해결하고자 하는 것이, 본 논문의 방법이다. 텍스트의 특징을 지니는 필드인 요약, 특허 내용 및 청구 범위를 각각 벡터로 표현하고, 숫자와 기호, 고유명사로 이루어진 필드들은 그 자체의 색인어를 기록하여, 이를 하나의 특허 문서에 대한 표현으로 삼는다.

벡터들과 색인의 조합으로 하나의 특허 문서를 저장하고, 특허 검색에 대한 질의가 들어올 때, 질의가 숫자/기호 필드(출원번호 등)에 대한 질의의 경우에는 기존의 불린 검색과 같은 식으로 동작하며, 내용에 대한 질의 (query by example, 문장질의)를 처리할 경우에는, 내용 필드들의 벡터와, 질의사이의 유사도를 계산하여 검색을 수행한다. 둘 모두가 포함된 질의의 경우에는, 불린 필드가 유사도를 비교할 제한조건으로 작동한다. 즉, 주어진 제한조건을 만족하는 문서들에 대해서만 유사도를 비교하여 결과를 내어놓는다.

### 2.3 IPC 분류를 이용한 가중치 부여

본 연구에서 벡터를 이루는 각 단어들에 대한 가중치부여는, TF-IDF 방법을 적용한다. 이때, 특허 문헌이 IPC 분류에 따라 출원 될 때부터 완전히 분류된 문서라는 점을 이용하여, IDF의 DF값을 보다 세밀하게 적용하려고 노력하였다. IPC (International Patent Classififation)는 국제특허분류라 하여 표2 에서와 같이 IPC는 Section, Class, Subclass, Main group, Subgroup으로 이어지는 계층 구조를 갖는다.

표2. IPC 의 구성

| Section | Class | Sub Class | / | Main Group | / | Sub Group |
|---|---|---|---|---|---|---|
| 영문자 1자리 | 숫자 2자리 | 영문자 1자리 | / | 숫자 3자리 | / | 숫자 2-4자리 |

IPC 분류표에서 section은 8개로 구성되어 있고 각 Section 은 Class로 구분되며 총 22개로 구성되어 있다.

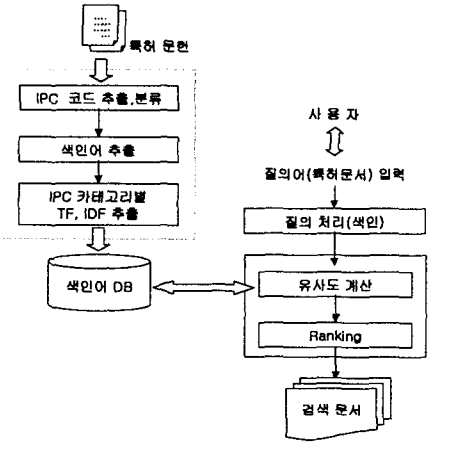
본 논문에서는 IPC의 5가지 계층구조 중 Section 아래의 Class 계층만을 이용하여 문서를 분류한다. 표3에서는 Section 과 class의 구조를 보여준다. 이렇게 22개의 분류만을 선택한 이유는, 이보다 더 하위 단계의 분류를 선택하는 경우에 더 좋은 결과를 기대할 수 있겠지만, 하나의 단어에 대해서 기록하게 되는 카테고리 빈도의 분류 개수가 너무 많아지기 때문이다. 수백만이 넘는 특허 문서에 대해서, 모든 색인되는 단어별로 이렇게 많은 숫자의 상이한 기록을 가지는 것이, 한 단계 더 내려가서 얻는 성능 개선만큼의 가치가 없다고 판단하였다.

표3. IPC의 상위 두 계층

| Section(A-H) | Class(22개) |
|---|---|
| A Section : 생활필수품 | A0 : 농업 |
| | A2 : 식료품; 담배 |
| | A4 : 개인용품 또는 가정용품 |
| | A6 : 건강,오락 |
| B Section : 처리조작 | B0 : 분리;혼합 |
| | B2 : 성형 |
| | B4 : 인쇄 |
| | B6 : 운수 |
| ... | ... |

## 3.특허 정보검색시스템의 구현

### 3.1 시스템 구성도




<그림1> 특허 검색 시스템의 구성도

그림1은 제안한 시스템의 전체적인 구성도이다. 먼저특허문서에서 IPC 코드를 추출하여 문서를 분류한다. 분류된 문서에서 색인어 추출기를 이용하여 색인어를 추출한후 분류된 카테고리별 단어빈도(TF)와 카테고리별 문헌빈도(DF),역 카테고리 빈도(IDF)를 구한다[6][7]. 이 값들을 벡터화 시켜 DB에 저장한다. 사용자가 질의를 하면 질의를 처리한 후 벡터화된 DB와 질의간에 유사도를 계산하고 순위화를 시킨다.

질의(q)와 문서(d)와의 유사도는 코사인유사도[8]를 이용한다.코사인 유사도는 아래 식(1)과 같다.

$$sim(d,q) = \frac{\sum_{k\in(q\cap d)} w_{kd} \cdot w_{kq}}{\sqrt{\sum_{k\in d} (w_{kd})^2} \sqrt{\sum_{k\in q} (w_{kq})^2}} \quad (1)$$

## 4.실험 및 평가

실험 대상은 2001년 국내 특허테이타 중 무작위로 추출한 11942건으로 하였다. 먼저 IPC 분류정보를 이용하여 문서를 8개의 Section과 하위 22개의 Class로 문서를 분류한 후 하나의 특허문서가 가지고 있는 여러 필드들 중 텍스트의 성격을 지니는 주요한 4가지필드(제목(TI),요약(AB),키워드(KW),청구범위(CL))추출하였다. 실험은 미리 답이 밝혀진 10개의 질의에 대해서 정확률과 재현율을 구하되, 방법의 차이에 따라 어떤 성능 변화가 보이는지를 밝히는 것에 주안점을 두었다.

실험은 크게 다음과 같이 나눌 수 있다. 먼저 TF-IDF 가중치 부여에서는 일반적인 IDF 방법과 IPC분류를 고려한 IDF 방법으로 나누어 실험하였다. 또한 문서를 이루는 필드 모두를 하나의 벡터로 보아 실험한 경우와, 주요한 텍스트인 요약과 청구범위를 각각 따로 벡터로 표현하고, 타 특허 문서의 해당 필드와의 유사도만을 비교한 결과를 구해보았다. 이 결과가 표 4와 표 5에 표현되어 있다.

표4.검색성능의 비교

| 검색모델 (임계치) | IPC Vector Model(0.2) | | IPC Vector Model(0.3) | | Vector Model(0.2) | | Vector Model(0.3) | |
|---|---|---|---|---|---|---|---|---|
| | 정확률 (%) | 재현률 (%) | 정확률 (%) | 재현률 (%) | 정확률 (%) | 재현률 (%) | 정확률 (%) | 재현률 (%) |
| 특허 문서 | 43.17 | 100 | 60.85 | 86.71 | 23.61 | 100 | 40.37 | 92.85 |

표5. 각 필드별 비교

| 검색모델 (임계치) | IPC Vector Model(0.2) | | IPC Vector Model(0.3) | | Vector Model(0.2) | | Vector Model(0.3) | |
|---|---|---|---|---|---|---|---|---|
| | 정확률 (%) | 재현률 (%) | 정확률 (%) | 재현률 (%) | 정확률 (%) | 재현률 (%) | 정확률 (%) | 재현률 (%) |
| AB | 41.81 | 92.85 | 66.81 | 89.28 | 29.76 | 96.42 | 38.05 | 85.17 |
| CL | 37.39 | 92.85 | 67.37 | 85.71 | 18.50 | 94.71 | 34.88 | 83.14 |

표4 의 결과는 특허 문서의 모든 필드를 하나의 벡터로 표현했을때의 결과이며, 표 5에서는 대표적인 내용 필드인 요약과 청구범위를 각각의 벡터로 표현하여, 그 필드간의 유사도를 특허건 간의 유사도로 처리했을 때의 결과이다. 일단 두 결과 모두, IPC분류 내에서 TF-IDF가중치를 구한 경우가 더 좋은 결과를 보인다. 표5의 결과에서는, 청구범위와 요약을 대상으로 하였을 때 특허 문헌의 요약 필드의 텍스트들이, 아마도 특허 문서를 더 잘 대표하는 것으로 보인다. 표4의 모든 필드를 하나로 모은 결과를 보면, 특허 문서의 일부인 요약필드와 청구 범위 필드만을 대상으로 한 결과보다 더 낮은 검색 성능을 보이기도 한다. 서론에서 살핀 바와 같이, 특허 문헌 내부의 여러 필드는 종종 상이한 성격의 텍스트로서, 이들이 하나의 벡터로 표현되면서 오히려 검색 결과에 지장을 줄 수도 있는 것으로 보인다.

## 5. 결론 및 향후과제

본 논문에서는 특허 문서의 특징을 살펴보고, 이를 반영하여 특허검색을 위한 벡터 스페이스 검색 모델을 제안하고 실험적으로 구현하였다. 특허 문헌 한 건을 이루는 요약, 청구 범위 등 각 필드를 단일한 하나의 텍스트로 보고 검색한 결과와, 요약 및 청구 범위 각각에 대해서 검색 결과를 구해보았다. 하나의 필드만을 선택할 경우, 요약 필드의 텍스트에 대한 유사도가, 높은 결과를 보였다. 각 필드별로 구할 수 있는 유사도 수치들을, 특허문헌 전체를 대표할 검색을 위한 하나의 수치척도로 통합하는 실험을 계속 수행할 계획이다.

텍스트의 성격을 지니는 필드들을 구분하여 하나 이상의 벡터로 취하고, 나머지는 색인어 자체를 포함해 문서의 표현으로 삼는 것이, 특허 검색 시스템의 구축에 가장 적절한 방법으로 보인다. 실험을 통하여, IPC분류 기준을 사용한 가중치가, 더 좋은 결과를 보임을 밝혔다.

실험에 사용된 테스트셋의 크기는 너무 작아, 실환경의 특허 문서 검색 시스템이 지니는 난점들이 나타나 있지 않다. 매년 추가되는 특허 문서는 너무나 많으며, 이를 고려해서 시간적으로 빠르며, 공간적으로 문서표현이 간결한 검색 시스템을 만들어야 한다. 또한 하나 이상의 벡터로 표현된 문서를, 어떻게 좋은 결과를 나타내는 검색과 순위 표시를 매겨줄 것인가에 대한 연구도 반드시 더 필요하다.

## 6. 참고문헌

[1]특허청. 특허청연보 1998, 서울: 특허청, 1998.

[2]WIPO, International Patent Classification : General Information(Geneva : WIPO 1994).

[3]권영숙,"특허문헌의 내용구조에 의한 특허정보 검색시스템 설계,"중앙대학교 문헌정보학과,2000.

[4]김해숙,"국제특허분류표의 분석적 고찰:Ranganathan의 문헌이론을 바탕으로," 계명대학교 도서관학과,1997.

[5]방용주,"특허정보 검색효율 증대 방안에 관한 연구," 연세대학교 산업대학원, 1987.

[6]T.Joachims,"A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," Proc. 14th Conf. on Machine Learning, Nashville TN,1997.

[7]G.Salton and C.Buckely,"Improving Retrieval Performance by Relevance Feedback," Journal of the American Society for Information Science, 41(4), 1990.

[8]Berthier Ribeiro-Neto,Ricardo Baeza-Yates,"Modern Information Retrieval," Addison-Wesley,1999.