한국산학기술학회논문지
Vol. 11, No. 4, pp. 1458-1467, 2010

# 특허 등록 예측을 위한 특허 문서 분석 방법

구정민[1], 박상성[1], 신영근[1], 정원교[1], 장동식[1*]
[1]고려대학교 정보경영공학부

# Analysis method of patent document to Forecast Patent Registration

**Jung-Min Koo[1], Sang-Sung Park[1], Young-Geun Shin[1], Won-Kyo Jung[1] and Dong-Sik Jang[1*]**
[1]**Division of Information Management Engineering, Korea University**

**요 약** 최근 지식재산권의 모방과 권리 침해는 국가 산업발전의 저해요소로 인식되고 있다. 많은 연구자들은 이러한 저해요소로 인하여 발생하는 막대한 손실을 막기 위해 지식재산권의 보호와 효율적 관리에 관한 연구를 다양하게 진행 중이다. 특히, 특허 등록 예측은 지식재산권 보호와 권리 주장을 위해 매우 중요한 연구이다. 본 연구는 텍스트 마이닝 기법을 이용한 특허문서 분석을 통하여 특허 등록 및 거절 여부를 예측하는 방법을 제안한다. 먼저 거절된 특허문서들의 단어 빈도수를 이용하여 데이터베이스를 생성한다. 그리고 생성한 데이터베이스와 다른 특허문서들을 비교하여 각 문서와 데이터베이스와의 유사한 정도를 판단하는 유사치를 도출한다. 본 논문에서는 특허 거절 기준 값을 선정하기 위하여 분할 군집화 알고리즘인 k-means 사용하였다. 그 결과로 거절된 특허 문서와 유사한 특허 문서는 거절될 가능성이 높다는 결론을 얻을 수 있었다. 실험을 위한 데이터는 현재 미국에 출원되어 있는 블루투스 기술, 태양전지 기술 그리고 디스플레이에 관한 특허 문서를 이용하였다.

**Abstract** Recently, imitation and infringement rights of an intellectual property are being recognized as impediments to nation's industrial growth. To prevent the huge loss which comes from theses impediments, many researchers are studying protection and efficient management of an intellectual property in various ways. Especially, the prediction of patent registration is very important part to protect and assert intellectual property rights. In this study, we propose the patent document analysis method by using text mining to predict whether the patent is registered or rejected. In the first instance, the proposed method builds the database by using the word frequencies of the rejected patent documents. And comparing the builded database with another patent documents draws the similarity value between each patent document and the database. In this study, we used k-means which is partitioning clustering algorithm to select criteria value of patent rejection. In result, we found conclusion that some patent which similar to rejected patent have strong possibility of rejection. We used U.S.A patent documents about bluetooth technology, solar battery technology and display technology for experiment data.

**Key Words :** Patent Forecast Text Mining

## 1. 서론

정보통신 기술과 인터넷의 발달은 개인이 소유하고 있는 다양한 정보와 지식을 쉽게 공유할 수 있게 되었다. 우리는 공유된 지식이나 정보를 이용하여 상품을 개발하고 판매하여 수익을 얻을 수 있다. 이러한 과정을 통해 수익을 창출할 수 있는 지식과 정보는 지식재산권이라고 하며 이를 법으로 보호하는 제도가 특허이다. 만약 특허를 등록하여 그 권리를 소유하면 그 기술을 사용하는 자로부터 로열티를 받을 수 있는 권리도 얻게 된다. 일례로

본 논문은 2010년도 두뇌한국 21사업에 의하여 지원되었음.
*교신저자 : 장동식(jang@korea.ac.kr)
접수일 10년 02월 05일 수정일 10년 04월 07일 게재확정일 10년 04월 09일

미국의 IBM의 경우 매년 라이선스료로 20억 달러의 추가적인 이익을 얻고 있다.[1] 이렇듯 특허를 출원하여 등록하는 것은 경제적인 이익을 얻을 수 있는 지적인 권리를 획득하는 것이라고 할 수 있다. 그러나 출원한 특허는 모두 등록되지는 않으며 등록되지 못하면 법으로 보호받지 못하므로 그 가치가 줄어든다. 또한 특허를 출원하는데 드는 비용이 적지 않기 때문에 경제적인 부담도 적지 않다. 따라서 등록되지 못할 특허를 출원하는 것은 경제적인 손실이라고 할 수 있다. 따라서 본 연구에서는 텍스트 마이닝을 이용하여 특허의 등록 및 거절 여부를 예측하기위한 특허 문서 분석 방법을 제안한다. 제안한 방법은 거절된 특허문서들의 단어 빈도수를 이용하여 데이터베이스를 생성한다. 그리고 생성한 데이터베이스와 다른 특허문서들을 비교하여 각 문서와 데이터베이스와의 유사한 정도를 판단하는 유사치를 도출한다. 이러한 특허 문서에 대한 유사치를 구하는 방법을 다른 특허 문서들에 적용하고 k-means를 이용한 실험을 하였다. 그 결과로 거절된 특허 문서와 유사한 특허 문서는 거절될 가능성이 높다는 결론을 얻을 수 있었다. 제안한 방법의 실험을 위한 데이터는 미국에 출원된 블루투스 기술과 태양전지 기술에 관한 특허 문서를 이용하였다.

본 논문은 2절에서 관련연구와 제안한 방법의 기초인 텍스트 마이닝에 대해서 살펴보고 3절에서 특허의 등록 요건에 대해서 알아본다. 4절에서는 제안하는 방법에 대해서 설명하며 5절에서는 실험을 통해 결과를 측정한다. 6절에서는 본 논문의 결론 및 향후 연구방향에 대해서 살펴본다.

## 2. 관련 연구

최근까지 특허에 관한 연구는 출원날짜, 출원인, 인용관계와 같은 구조화된 정보를 기반으로 한 특허 분석이 주요 과제였다.[2,3,4,5] 이러한 구조화된 데이터들은 서지적인 방법(bibliographic method)들을 이용하여 분석할 수 있다. 서지적인 방법들에는 데이터 마이닝 기술이나 OLAP (On-Line Analytical Processing)와 같이 잘 구성되어 있는 데이터베이스 관리 툴 등을 이용한 것들이 있다. 그리고 최근에는 특허 분석과 특허 맵을 구성하는 작업을 하기 위한 텍스트 마이닝 기술이 이슈화 되고 있으며, 텍스트 마이닝을 이용하여 특허 문서를 분석하고 그 특허 문서로부터 기술 트렌드를 찾기 위한 연구도 진행 중이다.[6-9]

아직까지 텍스트 마이닝을 이용한 특허 연구는 통계적인 분석을 하거나 인용관계, 특허 맵 생성, 그리고 기술트렌드 분석에 관한 것들이 대부분이며 특허의 등록 가능성에 관한 연구는 거의 이루어지지 않고 있다. 따라서 본 논문에서는 텍스트 마이닝 기법을 이용하여 특허 문서의 단어를 추출하고 이를 바탕으로 특허의 등록 가능성을 예측하는 연구를 하였다. 텍스트 마이닝이란 텍스트로 구성되어 있는 데이터에 데이터 마이닝 기법을 적용하는 것이라 할 수 있다. 텍스트 마이닝에서는 일반적으로 특징 벡터(feature vector)를 이용하는 기법을 사용한다. 이 기법은 특징 추출 과정을 통하여 텍스트에 대한 특징 벡터를 생성하는데 이 과정을 통해서 텍스트 문서에서 중요한 용어(term)를 인식하여 추출한다. 그리고 여기서 추출된 용어들은 특징 벡터를 구성하기 위해 단어의 원형(word)으로 변형되고 이 단어의 원형들은 문서를 요약하거나 분류 할 때 기초적인 정보로 사용된다. 여기서 특징(feature)의 중요도는 문서에서 단어가 나타나는 위치와 나타나는 횟수에 따른다. 즉 한 문서에서 어떤 단어의 빈도수가 높으면 중요도가 높다고 가정할 수 있지만, 반대로 그 단어의 빈도수가 낮으면 상대적으로 중요도가 낮다고 가정할 수 있다.[10,11]

본 논문에서는 단어의 빈도수가 높으면 그 문서에서 중요도가 높다고 가정하였다. 따라서 특허 문서에서 차지하는 비율이 높은 단어가 그 문서의 특징을 나타내는 핵심단어라고 가정하고 이 핵심단어들을 추출하여 특허 문서의 특징을 나타내는 값을 도출하는 방법을 제안하였다.

## 3. 특허의 등록 요건

특허가 등록이 되기 위해서는 그 발명이 특허로 인정받기 위한 요건을 갖추어야한다. 이러한 특허의 실체적 요건으로는 특허법상의 '발명'이면서 산업상 이용가능성, 진보성, 신규성이 있다. 특허법 상의 '발명'은 자연법칙을 이용한 기술적 사상의 창작으로서 고도한 것을 말한다. 산업상 이용가능성은 그 발명이 산업에 실제 활용될 수 있는 지를 의미한다.

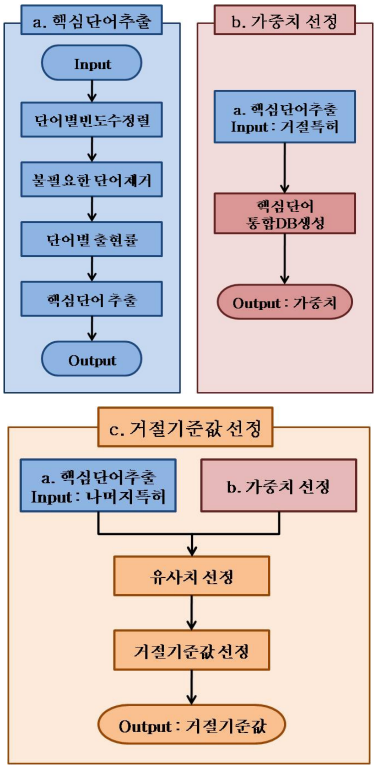
신규성은 발명의 내용인 기술적 사상이 종래의 기술적 지식, 선행기술에 비추어 알려져 있지 않은 새로운 것임을 의미한다. 신규성을 판단하기 위하여 시기적인 기준을 따르게 되는데 이 기준에 의해서 이미 기존에 존재하고 있는 기술에 대한 발명은 신규성을 상실하게 된다. 진보성은 신규성과 비슷한 개념으로 그 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 용이하게 발명할 수 있느냐에 의하여 판단한다. 이 진보성은 심사관의 지식과 경험에 따라서 주관적으로 판단된다.[12]

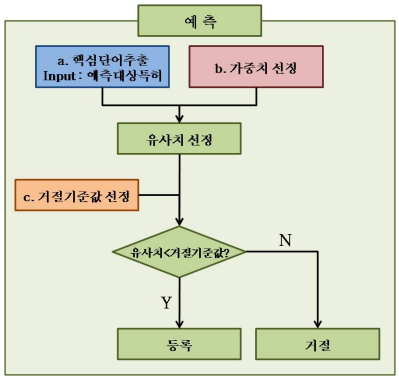
등록된 특허는 이러한 특허의 요건들을 모두 갖추고

있으며, 신규성에 의거하여 새로운 기술들에 관한 내용을 담고 있다고 할 수 있다. 새로운 기술에 관한 내용을 포함하고 있는 등록된 특허들은 대부분 유사한 기술에 관한 특허들이므로 각 문서 간에 중복되는 단어들의 사용이 많다. 그러므로 등록된 특허에서 단어의 빈도수를 사용해 특허의 특징치를 생성하는 것은 문제가 있다. 따라서 거절된 특허 문서들에서 거절이 되기 위한 특징을 찾아내는 작업이 필요하다. 따라서 본 연구에서는 거절된 특허를 기준으로 등록 및 거절 여부를 예측을 하기 위해 특허 문서의 특징치를 추출하는 방법을 제안하였다.

## 4. 제안된 방법

본 논문에서 제안하는 방법은 특허 문서가 저장되어 있는 웹상의 데이터베이스에서 검색식을 이용하여 특허 문서들을 추출한다. 그 후 텍스트 마이닝을 실행하여 각 단어들에 대한 가중치를 정하고 이를 바탕으로 등록가능성을 예측한다. 제안한 방법의 절차는 그림 1과 같이 구체적으로 나타낼 수 있다.




[그림 1] 제안된 방법의 순서도

제안한 방법은 크게 가중치 선정과정과 거절기준값 선정과정, 예측 과정으로 구성된다.

여기서 각 과정에 있는 단어를 빈도수 별로 정렬하는 단어별 빈도수 정렬 단계, 불필요한 단어를 제거하는 불필요한 단어제거 단계, 단어가 문서에서 나타나는 확률을 구하는 단어별 출현률 단계, 단어수의 평균보다 작은 출현률을 가진 단어를 제거하는 핵심단어 추출 단계는 동일하다. 첫 번째 과정인 가중치 선정과정에서는 거절된 특허 문서들에서 각 단어가 문서에서 나타나는 확률인 출현률과 단어수의 평균을 이용하여 문서의 주요 내용이 되는 핵심단어를 추출한다. 그 후 핵심단어를 통합한 핵심단어 통합 DB를 생성하고 각 단어에 대한 가중치를 선정한다. 거절기준값 선정과정에서는 핵심단어 통합 DB를 생성할 때 사용하지 않은 나머지 특허 문서들을 이용하여 핵심단어를 추출한다.

그 결과와 핵심단어 통합 DB의 가중치를 이용하여 유사치를 얻는다.

그 후 각 문서의 유사치와 각 특허 문서의 등록이나 거절에 대한 결과 여부를 값으로 산정하고 k-means 알고리즘을 이용한 분류를 실행하여 분류의 경계선 값인 거절기준값을 선정한다. 가중치 선정과정과 거절기준값 선정과정을 통해서 특허 문서의 특징값이 되는 유사치를 추출할 수 있다. 이 유사치를 바탕으로 예측 과정의 방법을 이용하여 특허의 등록 및 거절 여부를 예측할 수 있다. 예측 방법으로는 다양한 방법을 적용할 수 있을 것이며 본 연구에서는 k-means 알고리즘을 통해 얻은 분류 경계선 값을 거절 기준값으로 이용하였다. 예측 과정에서는 예측 대상 특허 문서를 이용하여 핵심단어를 추출한다. 여기서 핵심단어는 그 문서에서 나타나는 빈도가 전체단어들이 나타나는 빈도의 평균이상인 단어로 선정한다. 그 결과와 핵심단어 통합 DB의 가중치를 이용하여 유사치를 얻고 거절기준값 선정 단계에서 선정한 거절기

준값과 비교하여 등록 및 거절 여부를 판단한다. 제안한 방법의 각 과정에 앞선 전처리과정으로 특허 문서를 추출하는 과정이 필요하다. 특허 문서를 추출하는 과정에서는 등록 및 거절여부를 알고자 하는 특허의 분류에 해당하는 특허 문서들을 추출한다. 이를 위해서 특허 데이터베이스에 특허와 관련된 기술의 IPC 코드와 핵심 키워드를 입력한다. 본 논문에서는, 특허 검색사이트인 WIPS(www.wips.co.kr)에서 IPC코드와 키워드를 입력하였다. 예를 들어, 블루투스에 관한 특허를 추출하기 위하여 핵심 키워드인 bluetooth와 관련 IPC 코드인 H04B-007를 조합한 다음과 같은 검색식을 입력한다.

- (bluetooth*) AND (H04B-007*).IPC.

이후 추출한 특허문서들을 등록된 문서와 거절된 문서로 구분하여 각각 텍스트 파일의 형식으로 저장한다.

## 4.1 가중치 선정

### 4.1.1 단어별 빈도수 정렬

단어별 빈도수정렬 단계에서는 텍스트 파일로 변환한 특허 문서들을 단어의 빈도수 별로 정렬하는 프로그램을 이용하여 정렬하고, 각 문서별로 각각 저장한다. 이때, 1번 문서부터 m번 문서까지 있을 경우 각 특허 문서를 y로 표기한다. 즉 y=(1, 2, 3, 4, ... , m). 각 단어의 빈도수가 높은 순서대로 정렬하여 표 1과 같이 배열한다. 예를 들어 문서 y에서 n개의 단어가 있을 때 (i=1, 2, 3, 4, ... n), 빈도수가 가장 높은 단어를 배열의 1에, 두 번째로 높은 단어를 배열의 2에 둔다. 표 1은 블루투스 기술에 관한 한 특허 문서의 단어 빈도수별로 배열한 것이다.

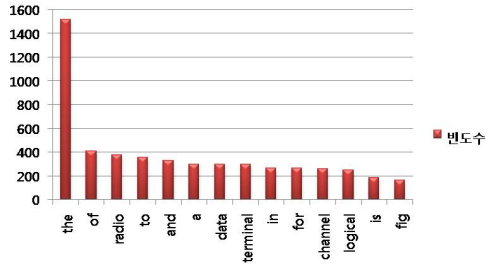
빈도수가 높을수록 그 문서에서 내용의 핵심이 되는 단어라고 할 수 있다.

**[표 1]** 단어의 빈도수별 배열

| i | 단어 | 빈도수 |
|---|---|---|
| 1 | the | 1465 |
| 2 | of | 416 |
| 3 | radio | 391 |
| ... | ... | ... |
| n | XXX | X |

### 4.1.2 불필요한 단어 제거

이 단계에서는 불필요한 단어를 제거한다. 그림 2는 단어의 빈도수별로 정렬한 예이다. 그림을 보면 특별한 의미가 없는 ‘the' 와 'a', 'to' 등의 단어가 빈도수의 상위권을 차지하고 있는 것을 알 수 있다. 따라서 단계 2의 불필요한 단어를 제거하는 과정이 필요하다. 이 과정에서 단어들 중, 기술적인 내용이 아닌 의미가 없는 단어들을 모두 제거한다. 예를 들면 ‘a, the’ 등의 관사나 ‘in, with’ 등과 같은 전치사 , 그리고 ‘invention, claim, fig’ 등과 같이 명세서를 작성할 때 필수적으로 들어가는 단어들이다. 표 2는 이 과정에서 제거할 단어들의 기준을 나타낸 것이다. 그리고 빈도수가 1인 단어도 모두 제거한다. 본 논문에서는 블루투스 기술에 관한 특허를 이용하였으므로 명세서에서 필수적으로 들어가서 의미가 없는 단어인 'bluetooth'도 제거하였다.

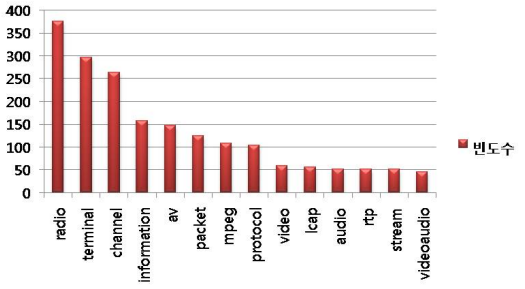



**[그림 2]** 단어의 빈도수별 정렬

**[표 2]** 단어 제거 기준의 예

| 제거기준 | 예 |
|---|---|
| 관사 | a, an, the 등 |
| 접속사 | and, but, so 등 |
| 전치사 | in, the, at, with 등 |
| 수사 | one, two, three 등 |
| 기타명사(대명사포함) | it, this, invention, bluetooth 등 |
| 기타 불필요한 단어 | same, all, easier 등 |

그림 3은 불필요한 단어들을 제거한 후, 빈도수별로 정렬한 것이다.




**[그림 3]** 불필요한 단어 제거

### 4.1.3 단어별 출현률 및 핵심단어 추출

이 단계에서는 각 단어가 문서에서 나타나는 비율인

단어별 출현률을 구하고 문서에서 핵심적인 내용을 구성하는 단어인 핵심단어를 추출한다. 출현률과 핵심단어를 추출하는 알고리즘은 다음과 같다.

**<핵심단어 추출 알고리즘>**

입력 : y는 특허 문서의 집합
i는 특허 문서 y에 있는 단어의 집합
i={1, 2, 3, ..., n}
출력 : 핵심단어 출현률 $P_{iy}*$
1. for(집합 y의 샘플 각각에 대해) {
2. for(i=1 to n) {
3. $P_{iy} = \frac{W_{iy}}{\sum_{i=1}^{n} W_{iy}}$;}
//$P_{iy}$는 문서y에 대한 단어 i의 출현률.
// $W_{iy}$는 문서 y에 있는 단어 i의 빈도수.
4. $C_y = \frac{1}{n}$;
// $C_y$는 문서 y의 핵심단어 추출 기준값.
5. if($P_{iy} < C_y$) {
6. remove $P_{iy}$;}
7. else {
8. $P_{iy}*=P_{iy}$;}
10. }

단계 3에서는 단어 i의 출현률 $P_{iy}$를 구하기 위해 각 단어의 빈도수를 전체단어들의 빈도수의 합으로 나눈다. 단계 4에서는 문서에서 핵심적인 내용을 구성하는 단어들을 추출하기 위한 핵심단어 추출 기준값을 구한다. 핵심단어 추출 기준값은 문서의 전체 단어수인 n의 평균으로 하였다. 평균 이상의 비중을 가지고 있는 단어들이 그 문서의 핵심적인 내용들을 가지고 있는 것들이다. 단계 6에서는 단계 4에서 구한 핵심단어 추출 기준값인 $C_y$를 이용하여 $P_{iy} < C_y$인 단어는 모두 제거한다. 그리고 문서 y에서 $P_{iy} < C_y$인 단어를 제거한 나머지 단어들의 출현률을 $P_{iy}*$로 표기한다.

#### 4.1.4 핵심단어통합DB 생성

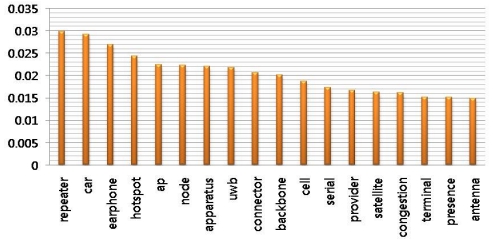
핵심단어통합DB 생성 단계에서는 핵심단어추출 단계를 거친 특허 문서들을 모아서 데이터베이스로 만들고 핵심단어 각각에 대한 가중치를 선정한다. 즉 거절된 특허 문서들에서, $P_{iy} < C_y$인 단어를 제거한 나머지 핵심단어들과 그 출현률인 $P_{iy}*$가 핵심단어통합DB를 구성하게 되는 것이다. 본 연구에서는 핵심단어통합DB에 있는 핵심 단어들의 출현률을 그 핵심단어의 가중치로 설정하고 $V_i$로 표기한다. 핵심단어통합DB의 가중치를 선정하는 알고리즘은 다음과 같다.

**<핵심단어통합DB 생성 알고리즘>**

입력 : y는 거절된 특허 문서의 집합
y={1, 2, 3, ..., m}
i는 특허 문서 y에 있는 단어의 집합
출력 : 가중치 $V_i$, 핵심단어통합 DB
$P_{iy}*$은 <핵심단어 추출 알고리즘>의 결과
1. for(y=1 to m) {
2. if(핵심단어가 중복될 경우) {
3. $V_i = \frac{\sum_{y=1}^{m} P_{iy}*}{R_i}$;}
// $V_i$는 단어 i의 가중치
//$R_i$는 중복되는 핵심 단어 i의 수
4. else {
5. $V_i$=$P_{iy}*$;}
6.}

출현률인 $P_{iy}*$값은 <핵심단어 추출 알고리즘>으로 구한 것이다. 핵심단어통합DB를 만들 때, 특허 문서들을 하나로 합치기 때문에 중복되는 단어들이 발생하게 된다. 단계 3의 과정을 통해서 이러한 중복되는 핵심단어들의 가중치를 설정한다. 만약에 핵심단어통합DB에서 중복되는 핵심단어들이 없을 경우 그 단어의 값이 가중치가 된다.

그림 4는 핵심단어통합DB생성 단계의 결과로서, 블루투스 기술에 관한 특허로 만든 핵심단어통합DB의 모습이다.




**[그림 4]** 데이터베이스

### 4.2 거절기준값 선정

거절기준값 선정 과정에서는 핵심단어통합 DB를 생성할 때 사용하지 않은 특허 문서들을 대상으로 <핵심단어추출 알고리즘>을 실행한 결과를 이용한다.

#### 4.2.1 유사치 선정 알고리즘

유사치를 선정하기 위한 알고리즘은 다음과 같다.

**<유사치 선정 알고리즘>**

입력 : y는 핵심단어통합DB 생성 단계에서 사용하지 않은 특허문서의 집합
i는 핵심단어통합DB와 y에서 중복되는 단어들의 집합 i={1, 2, 3, 4, ... p}

출력 : 유사치 $S_y$

$V_i$는 핵심단어통합DB에 있는 각 단어의 가중치

1. for(집합 y의 샘플 각각에 대해) {
2. <핵심단어추출 알고리즘> 수행;
3. for(i=1 to p) {
4. $S_y = \sum_{i=1}^{p} (V_i \times P_{iy}^*)$;}
   // $V_i$는 단어 i의 가중치
   //$S_y$는 문서 y의 유사치
5. }

유사치 $S_y$는 핵심단어통합DB를 생성하는데 사용하지 않은 특허문서들에 대하여 <핵심단어추출 알고리즘>을 수행한 결과인 $P_{iy}^*$와 핵심단어통합DB의 가중치 $V_i$와의 곱의 합으로 결정된다. 여기서 곱셈 연산은 핵심단어통합DB와 중복되는 단어들의 $V_i$값과 가중치인 $P_{iy}^*$값으로 수행한다.

| 핵심단어통합DB | | |
|---|---|---|
| i | 단어 | $V_i$ |
| 1 | radio | 0.008 |
| 2 | sniff | 0.009 |
| 3 | digest | 0.005 |
| 4 | stack | 0.003 |
| ... | ... | ... |

⇔

| 비교특허(y) | | |
|---|---|---|
| i | 단어 | $P_{iy}^*$ |
| 1 | radio | 0.029 |
| 2 | lcap | 0.004 |
| 3 | rtcp | 0.002 |
| 4 | stack | 0.001 |
| ... | ... | ... |

**[그림 5]** 유사치 비교의 예

본 연구의 블루투스 기술에 대한 유사치 비교 과정에서 핵심단어통합DB와 비교 특허 문서가 그림 5와 같이 나왔다. 이 경우 유사치는 다음과 같다. 핵심단어통합DB와 비교 특허에서 중복되는 단어는 'radio'와 'stack'이므로

$$\begin{aligned} S_y &= (V_2 \times P_{1y}^*) + (V_4 \times P_{3y}^*) \\ &= (0.008 \times 0.029) + (0.003 \times 0.001) \\ &= 0.000003 \end{aligned}$$

#### 4.2.2 거절기준값 선정

본 연구에서는 거절기준값을 선정하기 위하여 현재 가장 널리 쓰이는 분할 군집화 알고리즘인 k-means 알고리즘을 이용한다. k-means 알고리즘은 구현이 매우 쉬우며 수행이 매우 빠르다는 장점이 있다. 그리고 군집의 개수인 k를 직접 지정할 수 있으므로 본 연구에서 채택하였다.[13] 여기서 군집의 개수 k는 등록 군집과 거절 군집을 의미하는 2를 입력하며, 입력데이터는 각 특허 문서의 유사치인 $S_y$와 실제 등록, 거절 결과를 0, 1로 표현한 값이 된다. 출력데이터는 등록 군집과 거절 군집의 경계선이며, 이 경계선의 값이 거절기준값이 된다. 이 경계선의 값은 등록된 특허 문서들의 유사치와 거절된 특허 문서들의 유사치를 바탕으로 특허 문서들의 군집을 분류한 일종의 군집 분류 기준 값이라고 할 수 있다.

표 3은 k-means 에 입력하기 위한 본 연구의 블루투스 기술에 대한 데이터의 실제 예를 나타낸 것이다. 등록여부 값에서 등록은 1, 거절은 0으로 나타내었다.

**[표 3]** k-means 입력데이터

| | 문서1 | 문서2 | 문서3 | ... |
|---|---|---|---|---|
| 유사치 | 3.76091 | 8.84324 | 2.75568 | ... |
| 등록여부 | 1 | 0 | 1 | ... |

### 4.3 예측

예측 단계에서는 <핵심단어추출 알고리즘>과 <유사치 선정 알고리즘>을 거친 예측 대상 특허 문서와 거절기준 값 선정단계에서 구한 거절기준값을 이용하여 등록 가능성을 예측한다. 예측을 하기 위한 알고리즘은 다음과 같다.

**<예측 알고리즘>**

입력 : y는 예측 대상 특허 문서의 집합
i는 특허 문서 y에 있는 단어의 집합
거절기준값 선정단계에서 구한 등록기준 값 C
출력 : 등록 여부 판정
1. for (집합 y의 샘플 각각에 대해) {
2. for (집합 I의 샘플 각각에 대해) {
3. <핵심단어추출알고리즘> 수행;
4. <유사치선정알고리즘> 수행;
5. if(C>$S_y$) {
등록 판정;}
6. else if(C<$S_y$) {
거절 판정;}
7. else {
판정 불가;}
8. }
9. }

예측 대상 특허 문서의 등록 여부를 판단하기 위해서는 <핵심단어추출 알고리즘>과 <유사치 선정알고리즘>을 거쳐서 그 문서에 대한 유사치를 얻는다. 그리고 그 문서에 대한 유사치와 거절기준값을 비교하여 거절기준값보다 클 경우에는 거절, 작을 경우에는 등록될 가능성이 높다고 판단한다. 거절기준값이 같을 경우에는 등록여부의 판단이 불가능하다.

유사치는 거절된 특허들이 가지고 있는 특징 단어를 비교문서가 얼마나 가지고 있는지에 대한 정도라고 할 수 있다. 즉 특허문서의 유사치가 높으면 그 문서는 거절된 특허가 가지고 있는 거절되기 위한 특징을 많이 가지고 있다고 할 수 있다. 예를 들어 본 연구에서 블루투스 기술에 관한 특허 문서 중, 한 거절된 특허 문서 'Radio communication system' 에서 'audio', 'headset', 'earpiece', 'twoway' 라는 단어가 가장 많은 비율로 나타났다. 이 특허 문서의 유사치는 9.661로 나왔다. 그리고 한 등록된 특허 문서 'Data transfer method and radio terminal for executing transport layer protocol on radio network' 에서는 'terminal', 'channel', 'rtcp', 'identifier' 라는 단어가 가장 많은 비율로 나타났으며, 유사치는 2.887로 나왔다. 블루투스 기술의 전체 특허 문서에 대한 거절기준값은 7.211로 나왔다. 특허 문서의 유사치와 거절기준값을 비교했을 때 거절된 특허 문서는 거절기준값 이상으로 유사치를 나타내므로 거절로 판정이 되었고, 등록된 특허 문서는 거절기준값 이하의 유사치를 나타내므로 등록으로 판정되었다. 거절기준값 이상의 유사치라는 것은 거절된 특허에서 많이 나타나는 단어들의 특징들을 더 많이 가지고 있다고 판단할 수 있다.

## 5. 실험 및 결과

본 논문에서는 미국에 출원되어 있는 블루투스 기술에 관한 특허 문서를 이용하여 실험을 하였다. 2009년 1월에 공개되어 있는 특허문서를 대상으로 하였으며, 검색 결과로 전체 545개의 특허문서 중 등록된 특허 204개와 거절된 특허 57개를 추출하였다.

핵심단어통합DB를 만들기 위해서 거절된 특허 35개를 사용하였으며 거절기준값을 구하기 위해 140개의 데이터를 사용하였다. 그리고 나머지 86개의 데이터를 예측 테스트 데이터로 사용하였다. 거절기준값을 구하기 위해서 140개의 문서데이터에 대한 k-means 알고리즘 실험을 한 결과 거절기준값인 7.2를 얻을 수 있었다. k-means 알고리즘의 입력 데이터는 140개이며 군집의 개수인 k값은 2를 입력하였다.

**[표 4]** 블루투스 기술에 관한 실험 결과

| | 거절 | 등록 | 합계 |
|---|---|---|---|
| **사용데이터** | 57 | 204 | 261 |
| **DB생성** | 35 | . | 35 |
| **거절기준값생성** | 12 | 128 | 140 |
| **테스트 데이터** | 10 | 76 | 86 |
| **맞은 개수** | 8 | 56 | 64 |
| **틀린 개수** | 2 | 20 | 22 |
| **전체 정확도** | | **74.4%** | |

이 기준 값으로 테스트 데이터 86개에 대한 등록여부 판정을 실시하여 표 4와 같은 결과를 얻을 수 있었다. 전체 테스트 데이터 수 86개 중, 예측이 맞은 수는 64개이고 틀린 개수는 22개로 예측의 전체 정확도가 74.4%가 나왔다. 여기서 전체 정확도는 다음과 같은 식을 기준으로 평가하였다.

$$\text{전체 정확도} = \frac{\text{전체 맞은 데이터 수}}{\text{전체 테스트 데이터 수}}$$

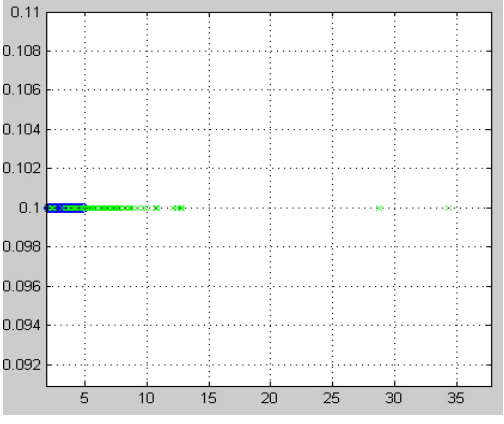
거절된 특허의 테스트데이터 10개 중 맞은 개수는 8개, 틀린 개수는 2개로 80%의 전체 정확도를 보였으며, 등록된 특허의 테스트데이터 76개 중 맞은 개수는 56개, 틀린 개수는 20개로 약 73.6%의 전체 정확도를 나타냈다. 거절된 특허가 등록된 특허에 비해 더 높은 전체 정확도를 나타내었으나 데이터 수가 상대적으로 적으므로

좀 더 정확한 전체 정확도를 평가하기 위해서는 더 많은 데이터가 필요하다고 판단하였다.

따라서 좀 더 많은 데이터에 대한 전체 정확도를 평가하기 위해 태양전지 기술에 대한 특허 문서들을 추출하였다. 추출한 특허 문서들 중 거절된 특허문서 100개로 핵심단어통합DB를 만들고 여기서 사용하지 않은 특허문서 400개를 이용하여 거절기준값인 4.6을 얻을 수 있었다. 그리고 테스트 데이터 200개로 전체 정확도의 판단 실험을 수행하여 표 5와 같은 결과를 얻을 수 있었다. 실험을 위하여 총 700개의 데이터를 사용하였다. 거절된 특허의 테스트 데이터 100개중 맞은 개수는 81개로 81%의 전체 정확도를 보였으며, 등록된 특허의 테스트 데이터 100개중 맞은 개수는 85개로 85%의 전체 정확도를 보였다. 등록된 특허가 좀 더 좋은 전체 정확도를 보였으나 차이가 크게 나지 않으므로 한쪽에 치우치지 않는 예측 결과를 얻을 수 있다고 판단된다. 그림 6은 테스트 데이터 200개의 데이터분포이다. 그림 6에서 거절기준값인 4.6의 오른쪽 편에 상대적으로 넓게 분포되어 있는 데이터가 거절된 특허들의 유사치이다. 거절된 특허의 데이터는 등록된 특허의 데이터보다 상대적으로 더 넓은 분포를 보였다. 등록된 특허의 데이터는 거절기준값인 4.6을 기준으로 좀 더 조밀한 분포를 보였다. 여기서 x축에 있는 숫자가 각 특허 데이터의 유사치를 의미한다.

[표 5] 태양전지기술에 관한 실험 결과

| | 거절 | 등록 | 합계 |
|---|---|---|---|
| 사용데이터 | 400 | 300 | 700 |
| DB생성 | 100 | . | 100 |
| 거절기준값생성 | 200 | 200 | 400 |
| 테스트데이터 | 100 | 100 | 200 |
| 맞은 개수 | 81 | 85 | 166 |
| 틀린 개수 | 19 | 15 | 34 |
| 전체 정확도 | | 83% | |

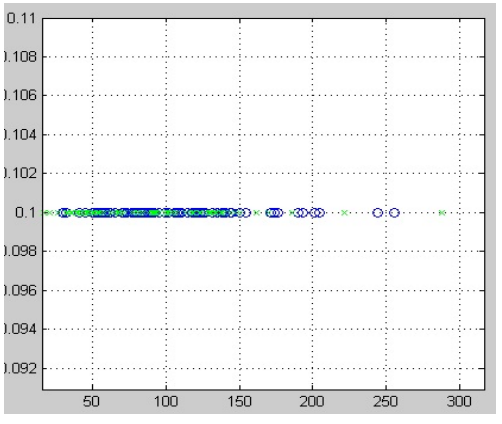



[그림 6] 태양전지기술의 데이터 분포

추가적으로 등록된 특허를 기준으로 특허의 등록여부를 예측하는 실험을 수행하였다. 동일한 조건을 위해 본 연구에서 제안한 동일한 알고리즘을 사용하였으며, DB 생성에 사용하는 등록 특허 문서의 개수도 동일하게 선정하였다. 태양전지기술에 관한 등록된 특허를 기준으로 하여 실험을 수행한 결과 표 6과 같은 결과를 얻을 수 있었다. 등록된 특허 문서 400개 중 100개를 DB 생성에 사용하였으며 나머지 거절기준값과 테스트 데이터에 사용한 개수는 거절된 특허 문서를 기준으로 한 이전의 실험과 동일하다. 실험 결과 거절된 특허의 테스트 데이터 100개 중 맞은 개수는 54개로 54%의 전체 정확도를 보였으며 등록된 특허의 테스트 데이터 100개 중 맞은 개수는 58개로 58%의 전체 정확도를 보였다. 전체 정확도는 56%로 예측결과가 등록, 거절의 두 개로 출력되는 것을 감안할 때 56%의 전체 정확도는 효율적인 예측 결과로 보기에 부적절하다고 할 수 있다. 그림 7은 등록된 특허를 기준으로 한 데이터 분포이다. 그림 7에서, 거절된 특허와 등록된 특허의 데이터가 혼합되어 넓게 분포되어 있으므로, 데이터의 정확한 분류가 거의 불가능한 것을 알 수 있다. 결과적으로 본 연구에서 제안한 알고리즘을 적용하여 등록된 특허를 기준으로 DB를 생성하고 예측을 하는 것은 불가능하다고 할 수 있다.

[표 6] 등록된 특허를 기준으로 한 실험 결과

| | 거절 | 등록 | 합계 |
|---|---|---|---|
| 사용데이터 | 300 | 400 | 700 |
| DB생성 | . | 100 | 100 |
| 거절기준값생성 | 200 | 200 | 400 |
| 테스트데이터 | 100 | 100 | 200 |
| 맞은 개수 | 54 | 58 | 112 |
| 틀린 개수 | 46 | 42 | 88 |
| 전체 정확도 | | 56% | |




[그림 7] 등록 특허를 기준으로 한 데이터 분포

본 연구의 신뢰성을 높이기 위해서 표 7과 같이 디스플레이 기술에 관한 특허 문서 데이터로 추가 실험을 수행하였다. 실험을 위하여 거절된 특허 350개, 등록된 특허 250개, 총 600개의 특허 문서를 추출하였다. 350개의 거절된 특허 데이터 중 90개를 DB생성에 사용하였고 나머지 중의 150개를 거절기준값 생성에 사용하였다. 그리고 남은 110개의 데이터를 이용하여 전체 정확도 측정 실험을 수행하였다. 마찬가지로 등록된 특허 데이터 250개 중 150개로 거절기준값 생성에 사용하였으며, 나머지 100개를 전체 정확도  측정을 위한 테스트 데이터로 사용하였다. 전체 정확도 측정 결과 거절된 특허의 데이터는 110개 중 82개를 맞춰 74.5%의 전체 정확도를 보였고, 등록된 특허의 데이터는 100개 중 74개를 맞춰 74%의 전체 정확도를 보였다. 결과적으로 전체 테스트 데이터 210개에 대하여 74.2%의 전체 정확도를 보였다.

**[표 7]** 디스플레이 기술에 관한 실험 결과

| | 거절 | 등록 | 합계 |
|---|---|---|---|
| **사용데이터** | 350 | 250 | 600 |
| **DB생성** | 90 | . | 90 |
| **거절기준값생성** | 150 | 150 | 300 |
| **테스트데이터** | 110 | 100 | 210 |
| **맞은 개수** | 82 | 74 | 156 |
| **틀린 개수** | 28 | 26 | 54 |
| **전체 정확도** | | **74.2%** | |

## 6. 결론

본 논문에서는 텍스트 마이닝을 이용하여 특허의 등록 및 거절 여부를 예측하기 위한 특징을 추출하는 방법을 제안하였다. 제안된 방법은 거절된 특허 문서가 거절이 되기 위한 특징을 가지고 있다고 가정하여 등록 및 거절 여부의 기준으로 하였다. 제안된 방법에서는 텍스트 마이닝을 통해 거절된 특허 문서들의 단어들을 빈도수 별로 추출하여, 핵심단어통합DB로 만들고 다른 특허문서 데이터들과 비교하여 거절기준값을 설정하였다. 이 거절기준값을 바탕으로 거절된 특허들로 구성된 핵심단어통합DB와 중복되는 단어가 많을수록 거절이 되기 위한 특징을 많이 가지고 있다고 판단하였다. 그리고 새로운 특허 문서 데이터를 이용한 테스트를 통하여 제안한 방법의 성능을 평가하였다. 이와 같이 특허 등록 및 거절 여부를 예측하는 방법은 새로운 기술을 개발함과 동시에 상품을 제조함으로써 수익을 얻으려는 기업에게 그 기술의 투자 가치를 미리 알 수 있다는 장점이 있다. 또 등록될 가능성이 낮은 특허문서를 출원함으로써 특허의 출원 비용을 낭비하게 되는 경제적인 손실을 줄일 수 있을 것으로 기대된다.

향후 과제로는 다른 분야의 특허 문서 데이터에 대한 성능을 평가하여 적용가능성을 검증하고 다양한 데이터 마이닝 알고리즘을 적용하여 예측 결과를 비교하는 방법을 연구해야 하겠다.

## 참고문헌


[1] http://en.wikipedia.org/wiki/Software_patents

[2] Archibugi. D. and Pianta. M, “ Measuring technological change through patents and innovation survey”, Technovation, Vol.16, No.9, pp.451 – 468, 1996.

[3] Be’de’carrax. C. and Huot. C, "A new methodology for systematic exploitation of technology databases", Information Processing & Management, Vol.30, No.3, pp.407 – 418, 1994.

[4] Ernst. H., "Use of patent data for technological forecasting: the diffusion of CNC-technology in the machine tool industry", Small Business Economics, Vol9, No.4, pp.361 – 381, 1997.

[5] Lai. K.-K. and Wu. S.-J, "Using the patent co-citation approach to establish a new patent classification system", Information Processing & Management, Vol.41, No.2, pp.313 – 330, 2005.

[6] Fattori. M, Pedrazzi. G. and Turra. R. "Text mining applied to patent mapping: a practical business case", World Patent Information, Vol.25, No.4, pp.335 – 342, 2003.

[7] Lent. B, Agrawal. R. and Srikant. R, "Discovering trends in text databases", In Proceedings of international conference on knowledge discovery and data mining, 1997.

[8] B. G. Yoon and Y. T. Park, "A text-mining-based patent network: Analytical tool for high-technology trend", Journal of High Technology Management Research Vol.15, No.1, pp.37 – 50, 2004.

[9] Y. S. Tian, Y. H. Kim, Y. J. Jeong, J. H. Ryu, and S. H. Myaeng, "A Language Model and Clue based Machine Learning Method for Discovering Technology Trends from Patent Text", Journal of Korean Institute of Information Scientists and Engineers, Vol 36, No 5, pp.420-429, 2009.

[10] Clifton. C. and Cooley. R, "TopCat: Data Mining for Topic Identification in a Text Corpus", Proceedings of the Third European Conference of Principles and

Practice of Knowledge Discovery in Databases, 1999.
[11] Yang. Y, "An Evaluation of Statistical Approaches to Text Categorization", Journal of Information Retrieval, Vol.1, No.1-2, pp.69-90, 1999.
[12] Korea Intellectual Property Office, *Patent and Information Analysis*, Korea Intellectual Property Office, 2007.
[13] 오일석, *Pattern Recognition*, 교보문고, 2008.

**구 정 민(Jung-Min Koo)** **[준회원]**

- 2008년 2월 : 대구가톨릭대학교 컴퓨터정보통신공학부 (공학사)
- 2008년 9월 ~ 현재 : 고려대학교 정보경영공학부 석사과정

<관심분야>
특허 정보 분석, 데이터 마이닝

**박 상 성(Sang-Sung Park)** **[정회원]**

- 2006년 2월 : 고려대학교 산업시스템공학과 (공학박사)
- 2006년 5월 ~ 현재 : 고려대학교 BK21 사업단 연구교수

<관심분야>
컴퓨터 비전, 패턴인식, 전문가시스템응용, 지식관리

**신 영 근(Young-Geun Shin)** **[정회원]**

- 2005년 2월 : 고려대학교 산업시스템정보공학과 (공학사)
- 2005년 9월 ~ 현재 : 고려대학교 정보경영공학부 석 박사 통합과정

<관심분야>
패턴인식, 스케줄링, 인공지능

**정 원 교(Won-Kyo Jung)** **[정회원]**

- 2007년 2월 : 경희대학교 산업공학과 (공학사)
- 2007년 3월 ~ 2009년 2월 : 고려대학교 정보경영공학부 석사
- 2009년 3월 ~ 현재 : 고려대학교 정보경영공학부 박사과정

<관심분야>
객체지향응용, 프레임웍, 정보시스템

**장 동 식(Dong-Sik Jang)** **[정회원]**

- 1979년 2월 : 고려대학교 산업공학과 (공학사)
- 1985년 6월 : 텍사스 주립대학 산업공학과 (공학석사)
- 1988년 12월 : 텍사스 A&M 산업공학과 (공학박사)
- 1989년 3월 ~ 현재 : 고려대학교 정보경영공학부 교수

<관심분야>
Computer Vision, 최적화이론, 컴퓨터 알고리즘