텍스트 마이닝을 이용한 특허 등록 예측에 관한 연구

구정민*, 박상성*, 신영근*, 정원교*, 장동식*
*고려대학교 정보경영공학과
e-mail: leonag@korea.ac.kr

A Study on Prediction of Patent Registration using Text Mining

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

Abstract

Recently, as importance of knowledge property right is rising, a patent is being issue. A patent is exclusive rights of knowledge or technique, and it must be registered for approval of rights. Therefore, prediction of patent registration can be important information for company or individuals which gain profit using a patent. In this paper, we proposed a method for prediction of patent registration using text mining and a algorithm for constructing database.

1. 서 론

최근 정보통신 기술의 발달로 지식과 정보의 교류가 활발히 진행됨에 따라 이들이 새로운 가치 창출과 경쟁력의 원동력이 되고 있다.

이것은 지식과 정보의 소유자가 시장을 지배하는 이른바 지식기반사회가 도래한 것을 의미 한다. 지식과 정보를 소유한 기업과 개인은 이것을 이용하여 직접 제품을 생산하여 판매함으로써 이익을 얻을 수 있을 뿐 아니라 지식과 정보 자체를 생산자에게 대여함으로써 그에 대한 사용료를 수익으로 얻을 수 있게 된다. 이러한 새로운 시장 환경 조성을 가능하게 한 대표적인 예가 지식재산권 중 특허라 할 수 있다.

특허는 해당 특허가 영향을 미치는 국가 안에서 그 지식과 정보에 대한 권리를 독점적으로 기업이나 개인이 소유하게 된다. 또 다른 기업이나 개인에게 어떤 대가를 받고 그 권리를 양도할 수도 있으므로 그 자체가 수익 수단이라고 할 수 있다. 이렇게 지식과 정보가 수익을 얻을 수 있는 권리를 갖기 위해서는 먼저 특허를 출원한 후 등록을 하여야 한다. 하지만 출원한 특허가 모두 등록되는 것은 아니다. 출원된 특허는 심사관의 엄격한 심사를 통해서 특허 등록 요건인 신규성, 진보성, 산업상 이용가능성이 충족되어야만 등록될 수 있다. 또, 특허는 출원 후 1년 6개월 뒤 일반인에게 자동으로 공개되며, 심사 후 등록되지 못한 특허는 누구나 사용할 수 있게 되므로 그 특허를 출원한 사람에게는 아무런 가치가 없게 되며 출원 시 소요되는 적지 않은 비용으로 인하여 막대한 경제적인 손실을 입을 수도 있다. 그러므로 특허의 등록 여부를 미리 아는 것은 특허를 이용하여 수익을 얻으려는 기업이나 개인에게 중요한 정보가 될 수도 있다. 그러나 특허의 등록 여부는 심사관의 주관적인 판단에 의하여 결정되므로 예측하기가 쉽지 않다.

따라서 본 연구에서는 텍스트 마이닝을 이용하여 특허의 등록가능성을 예측하는 방법과 데이터베이스를 만들기 위한 알고리즘을 제안하였다. 제안한 방법은 웹상에 있는 특허 데이터베이스에서 특허문서를 검색하여 추출하고 텍스트 마이닝 기법을 이용하여

빈도수별로 정렬하여 데이터베이스를 만들고 등록 여부를 알고자 하는 특허 문서의 등록 가능성을 예측하는 것이다. 본 연구에서는 미국에 출원된 블루투스 기술에 관한 특허를 이용하였다.

본 논문은 2절에서 관련연구와 제안한 방법에서 사용한 텍스트 마이닝 기법에 대해서 살펴보고 3절에서는 본 연구에서 제안하는 등록 가능성 예측 방법과 데이터베이스를 만들기 위한 알고리즘에 대해서 설명한다. 그리고 4절에서는 본 논문의 결론 및 향후 연구방향에 대해서 살펴본다.

2. 관련 연구

최근까지 텍스트 마이닝 기술을 이용한 특허 분석 연구는 계속 진행되고 있다. 예를 들면 텍스트 마이닝을 이용하여 특허 분석에 필요한 문자 분할, 요약 추출, 특징 선택, 용어 결합, 클러스터 생성, 정보 매핑을 하거나 기술 중심 인덱스, 기술 주기 인덱스, 기술 키워드 클러스터 등을 구성하여 인용 분석을 위한 특허 네트워크를 만드는 연구가 있다.[1][2] 그리고 키워드 기반의 특허 맵을 만들고 특허를 분류하는 연구도 진행 중이다.[3][4]

그러나 아직까지 텍스트 마이닝을 이용한 특허 연구는 특허를 분류하거나, 인용관계, 특허 맵 생성에 관한 것들이 대부분이며 등록 가능성을 예측하는 것에 관한 연구는 거의 이루어지고 있지 않다. 따라서 본 논문에서는 특허의 등록 가능성을 예측하기 위하여 텍스트 마이닝 기법을 바탕으로 특허 문서에 있는 단어의 빈도수를 이용한 방법을 제안하였다.

본 연구에서 사용한 텍스트 마이닝 기법은 데이터 마이닝 기법 중 하나로써 텍스트로 구성되어 있는 구조화되지 않은 문서의 문장이나 단어들의 조합을 처리하여 구조화된 데이터를 생성하고, 그 데이터를 이용하여 새로운 지식이나 패턴과 같은 특징을 추출하거나 텍스트를 분류하는 것이다. [5][6][7]

일반적으로 텍스트 마이닝에서는 특징 벡터를 이용한다. 특징 벡터(feature vector)는 텍스트 문서에 대한 특징 추출 과정에서 생성된다. 이 과정에서 텍스트 문서에서 인식하여 추출된 중요한 용어(term)들은 단어의 원형(word)으로 변형되어 특징 벡터를 구성하게 된다. 이 단어의 원형(word)들은 문서를 요약하거나 분류할 때 기초적인 정보로 사용된다. 이때, 특징 벡터에서 특징(feature)의 중요도는 단어가 문서에서 나타나는 위치와 횟수에 따른다. 예를 들면, 어떤 문서에서 한 단어의 빈도수가 높을 때, 그 단어의 중요도가 높으며, 반대로 빈도수가 낮으면 그 단어의 중요도가 낮다고 할 수 있다. [8][9]

3. 제안된 방법

본 논문에서 제안한 방법은 웹에 있는 특허 데이터베이스에서 특허문서들을 추출하고, 이것들 중 등록된 특허 문서들을 대상으로 텍스트 마이닝을 실행하여 데이터베이스를 만든다. 그리고 이 데이터베이스를 이용하여 등록 여부를 알고자 하는 특허 문서의 등록 가능성을 예측한다.

[그림 2]는 제안한 방법의 절차를 간략하게 나타낸 것이다.

[그림 2] 제안된 방법의 순서도

본 논문에서 제안한 방법의 상세한 절차는 다음과 같다.

(1) 특허 DB에 검색식 입력

등록 가능성을 예측하려는 특허와 관련된 기술의 특허 문서들을 추출한다. 웹에 있는 특허 데이터베이스에 관련 기술의 검색어를 입력하여 특허 문서들을 검색한다. 본 논문에서는 미국에 출원된 블루투스 기술에 관한 특허를 대상으로 하였으며 WIPS(www.wips.co.kr)를 이용하였다. 그 검색식은 다음과 같다.

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

(2) 특허 문서 추출

특허 데이터베이스에서 검색하여 결과로 출력된 특허 문서들을 다운로드하여 저장한다. 저장한 특허 문서들은 텍스트를 추출할 수 있게 HTML과 같은 텍스트 파일의 형식으로 변환하여 각 문서별로 저장한다.

(3) 프로그램을 이용한 텍스트 마이닝

텍스트 마이닝 프로그램을 이용하여 저장한 텍스트 파일에서 단어들을 추출하여 빈도수가 높은 순서대로 정렬한다.

(4) 문서별로 핵심단어 추출

추출한 단어들이 각 문서에서 나타나는 비율을 구하고 불필요한 단어들을 제거한다. 불필요한 단어 제거 기준은 [표 2]와 같다.

[표 2] 불필요한 단어 제거 기준

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

추출한 단어들이 나타나는 비율을 이용하여 핵심단어를 추출하기 위한 기준값을 구한다. 이 기준값은 추출한 단어들이 나타나는 비율의 평균이며 기준값보다 낮은 비율의 단어들은 모두 제거한다. 여기서 남은 단어들이 핵심단어이다. 핵심단어는 반복되는 횟수가 많은 단어들로써, 그 특허문서의 특징적인 기술에 대한 내용을 설명하기 위해 사용하는 것들이다.

(5) 핵심 단어를 통합한 DB 생성

등록된 특허 문서들의 핵심단어들을 통합하여 데이터베이스를 만든다. 본 연구에서는 등록된 특허 문서를 특허의 등록 가능성을 예측하기 위한 기준으로 삼았다.

(6) 등록 가능성 예측

단계(5)에서 만든 데이터베이스를 이용하여, 등록여부를 알고자하는 특허 문서의 등록 가능성을 예측한다.

단계(5)의 데이터베이스는 다음과 같은 알고리즘을 이용하여 만들었다.

y : 문서의 집합 y={1, 2, 3, ...m}

i : 문서 y의 단어의 집합 i={1, 2, 3, ..., }

1. for(y=1 to m) {

2. $P_{iy} = \frac{W_{iy}}{\sum_{i=1}^{n} W_{iy}}$ (i=1, 2, 3, ... n일 경우)

W_{iy}는 문서 y에서 단어 i가 나타난 빈도수

3. i=1에서 n까지의 단어에 P_{iy}값을 대입.

4. $C_y = \frac{\sum_{i=1}^{n} P_{iy}}{n}$ (i=1, 2, 3, ... n일 경우)

5. $P_{iy} < C_y$인 단어는 모두 제거

6. } 남은 단어를 모두 통합.

제안된 알고리즘으로 만든 데이터베이스를 이용하여 20개의 특허 문서를 대상으로 [표 3]과 같은 결과를 얻었다.

[표 3] 예측 테스트 결과

전체데이터 수	맞은 개수	틀린 개수	예측률
20	15	5	75%

20개의 특허 문서 중 예측이 맞은 개수는 15개이고 틀린 개수는 5개로 75%의 예측률을 보임을 알 수 있었다. 추후 예측률을 높이기 위한 연구가 계속 진행되어야 할 것이다.

4. 결 론

본 논문에서는 텍스트 마이닝을 이용하여 특허의 등록 가능성을 예측하는 방법과 데이터베이스를 만들기 위한 알고리즘을 제안하였다. 제안된 방법은 다음과 같다. 웹에 있는 특허데이터베이스에서 특허 문서를 추출하고 빈도수별로 정렬하여 데이터베이스를 만들었다. 이 데이터베이스를 이용하여 등록 여부를 알고자하는 특허 문서의 등록 가능성을 예측하였다. 이때 데이터베이스는 등록된 특허 문서들을 대상으로 본 논문에서 제안한 알고리즘을 이용하여 만들었다. 그리고 40개의 특허문서를 이용하여 실험을 하고 그 결과를 얻었다.

특허의 등록 가능성을 예측하는 방법은 특허를 이용하여 수익을 얻으려는 기업과 개인에게 그 특허의 가치를 미리 알 수 있는 정보를 제공해주며, 등록될 가능성이 낮은 특허를 출원함으로써 입게 되는 경제적 손실을 줄일 수 있을 것으로 기대된다.

향후 과제로 좀 더 개선된 방법을 이용하여 예측

률을 높이고, 특허의 등록 요건인 신규성과 진보성을 반영한 예측 방법도 연구해야 할 것이다.

감사의 글

- ◆ 이 논문은 2009년도 두뇌한국 21사업에 의하여 지원되었음.
- ◆ 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음.
  (IITA-2008-(C1090-0801-0025))
- ◆ This work was supported by the IT R&D program of MIC/IITA [2007-S019-02] (Development of Digital Forensic System for Information Transparency).

참 고 문 헌

[1] Yuen-Hsien Tseng, Chi-Jen Lin, Yu-I Lin, Text mining techniques for patent analysis, Information Processing and Management 43, 2007, 1216 - 1247

[2] B.G. Yoon, Y.T. Park, A text-mining-based patent network: Analytical tool for high-technology trend, Journal of High Technology Management Research 15, 2004, 37 - 50

[3] S.J Lee, B.G. Yoon, Y.T. Park, An approach to discovering new technology opportunities: Keyword-based patent map approach, Technovation, In Press, Corrected Proof, Available online 20 December 2008

[4] Michele Fattori, Giorgio Pedrazzi, Roberta Turra, Text mining applied to patent mapping: a practical business case, World Patent Information, Volume 25, Issue 4, December 2003, Pages 335-342

[5] A-H Tan, "Predictive Self-Organizing Networks for Text Categorization", in Proceedings, Fifth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'01), Hong Kong, pp. 66-77, 2001.

[6] Lakshmi V., A-H Tan, and C-L Tan, Web Structure Analysis for Information Mining, Accepted by ICDAR'01 Workshop on Web Document Analysis, Seattle, September 10-13, 2001.

[7] Mooney J., "Using Information Extraction to Aid the Discovery of Prediction Rules from Text", in Proceedings of the Sixth ACM SIGKDD International Conference on KDD Workshop on Text Mining, pp 51-58, Boston, MA, August, 2000.

[8] Clifton C. and Cooley R., TopCat: Data Mining for Topic Identification in a Text Corpus, in Proceedings of the Third European Conference of Principles and Practice of Knowledge Discovery in Databases, Prague, Czech Republic, 1999.

[9] Yang Y., An Evaluation of Statistical Approaches to Text Categorization Journal of Information Retrieval, 1999.