J. Korea Saf. Manag. Sci. Vol. 15 No. 4 December 2013
http://dx.doi.org/10.12812/ksms.2013.15.4.289

ISSN 1229-6783(Print)
ISSN 2288-1484(Online)

빅데이터 분석 도구 R을 활용한 효율적인 특허 검색에 관한 연구

장 청 윤* · 장 정 환* · 김 석 주* · 이 현 근** · 이 창 호*
*인하대학교 산업공학과
**엘비세미콘

A study on the efficient patent search process using big data analysis tool R

Jing-Lun Zhang* · Jung-Hwan Jang* · Suk-Ju Kim* · Hyun-Keun Lee** · Chang-Ho Lee*
*Department of Industrial Engineering, INHA University
**LB Semicon

Abstract

Due to sudden transition to intellectual society corresponding with fast technology progress, companies and nations need to focus on development and guarantee of intellectual property. The possession of intellectual property has been the important factor of competition power. In this paper we developed the efficient patent search process with big data analysis tool R. This patent search process consists of 5 steps. We result that at first this process obtain the core patent search key words and search the target patents through search formula using the combination of above patent search key words.

Key words : Patent Search Process, R, Target Patent

1. 서 론

디지털 기술의 발달로 세계가 정보 및 지식이 주도하는 사회로 급변함에 따라 정보 과학기술이 경제 및 사회에 미치는 영향이 지대하게 되었다. 정보사회에 맞는 균형 있는 기술 발전과 창조적 활동을 촉진시키기 위해서는 지식 재산권의 발전이 불가결하며, 각 기업 및 국가들은 그들의 강화력을 키우기 위해 지식재산권에 대한 중요성을 강조하고 있다[2]. 또한 지적재산권은 시장에서 독점적 지위 확보가 가능하고, 특허와 관련한 분쟁을 예방할 수 있으며, 막대한 기술개발 투자비를 회수할 수 있는 확실한 수단이며, 확보된 권리를 바탕으로 추가 응용 기술개발이 가능하다는 이유 등으로 중요성이 강조되고 있다[4]. 지적재산권에 대한 중요성이 강조되면서 이러한 지적재산권을 이용하는 사업군이 있는데 이를 특허전문관리회사 또는 특허괴물(patent troll)이라 한다. 특허전문관리회사는 제품을 제조하거나 판매하지 않고 특허권 또는 지식재산권만을 집중적으로 보유함으로써 로열티 수입으로 이익을 창출하는 회사이다[8]. 2012년, 우리나라의 특허관리전문회사는 2,327개로 이 분야에 대한 사업이 활발히 되고 있는 것을 확인할 수 있다[3].

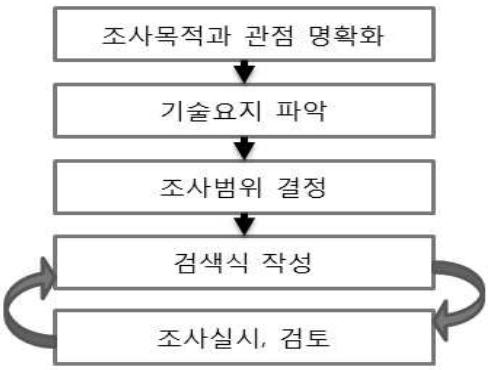
이와 같이 지식재산권의 중요성의 강조를 통해 현실에서 지식재산권의 확보는 기업의 경쟁력을 좌우하는 요소라 할 수 있다. 지식 재산권 확보를 위해서는 보다 정확하고 효율적인 특허 정보 조사가 필요하다. 따라서 본 논문에서는 빅데이터 분석 도구인 R을 이용하여 빠른 시간 안에 사용자가 목적으로 하고 있는 특허검색 결과를 효율적으로 도출할 수 있는 검색어 추출에 관한 연구를 진행하고자 한다.

† Corresponding Author : Chang-Ho Lee, Industrial Engineering, INHA UNIVERSITY, 100, inha-ro, Nam-gu, Incheon, M · P : 010-3761-2995, E-mail: lch5601@inha.ac
Received October 20, 2013; Revision Received December 5, 2013; Accepted December 5, 2013.

2. 이론적 배경

2.1 특허 검색의 개념

특허 정보 조사는 특허 정보의 특성을 활용하여 자신의 목적에 맞도록 특허 자료를 검색하는 것을 의미한다. 이러한 특허 정보 조사를 통해 관련 기술분야에서 기술개발의 흐름을 파악할 수 있고, 장래기술의 예측이나 선행기술조사로 중복 연구방지, 아이디어 입수, 기술동향을 파악할 수 있다[6]. 일반적으로 특허 검색을 위한 과정은 [Figure 1]과 같다.



[Figure 1] Patent search process

2.2 TF-IDF 알고리즘

TF-IDF(Term Frequency-Inverse Document Frequency)모델은 정보검색 및 텍스트마이닝을 위해 그리고 문서 내부의 단어 간 상대적 중요도를 평가하기 위해 문서의 표현방식으로서 만들어졌다. TF-IDF 모델은 벡터 공간모델(Vector Space Model) 기반 정보검색을 위해서 문서로 표현하는 원리를 사용하고 있다. TF-IDF값은 TF와 IDF를 곱한 값이다. TF값은 한 문서 내에서 특정 단어가 출현한 빈도수를 의미하며 문서 내부의 단어 출현 빈도를 모든 단어의 총 출현 회수로 나누어 정규화한 형태이다. IDF값은 문서 집합에 포함되어 있는 문서 수를 특정 단어가 나타난 문서의 수로 나눈 것이다[7]. 이를 수식으로 확인하면 <Table 1>과 같이 표현된다.

TF값이 크다는 것은 주어진 단어가 문서 내에서 많이 출현할수록 상대적으로 더 중요하다는 의미를 가지고 있으며, IDF값이 큰 단어는 문서 내에서 주요 의미를 가지는 단어로 분별이 된다. R 프로그램에서는 tm 패키지를 통해 이러한 기능을 구현할 수 있다.

<Table 1> TF-IDF algorithm

TF값	$n_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: 단어 ti가 문서 dj에서 출현한 회수 $\sum_k n_{k,j}$: 문서 dj에서 모든 단어가 출현한 회수
IDF값	$idf_i = \log \frac{\lvert D \rvert}{\lvert d_j \rvert t_j \in d_j \rvert}$ $\lvert D \rvert$: 문서집합에 포함되어 있는 문서의 수 $\lvert d_j \rvert t_j \in d_j \rvert$: 단어 t_j가 등장하는 문서의 수
TF-IDF 가중치	$TFIDF_{i,j} = tf_{i,j} \times idf_i$

2.3 Apriori 알고리즘

연관규칙은 데이터마이닝의 여러 기법들 중 하나로써 개인화 추천서비스에 널리 활용되고 있으며 장바구니 분석이라고도 불리는 것으로써 항목집합들로 이루어진 데이터베이스에서 항목들의 동시출현성향에 대한 관계성을 표현한다. 즉, 조건부 확률로써 “사건 A가 일어났을 때, 사건 B가 일어나는 것”을 나타내며, 추천시스템에서는 “임의의 고객이 조건 A를 만족할 경우, 조건 B를 만족한다.”를 의미한다. 여기서 조건 A는 고객의 나이, 성별 등의 특징이 될 수 있으며 또는 이용자가 선택한 항목들을 나타낼 수 있다. 그리고 조건 B는 이용자에게 추천할 항목으로 정의된다. 연관규칙 추출의 대상인 항목집합은 트랜잭션이라고 정의된다[5].

연관 규칙 마이닝 알고리즘 중 하나인 Apriori는 구매하는 물품들의 집합인 트랜잭션으로부터 연관 규칙을 마이닝한다. 연관 규칙은 두 단계를 통하여 구성된다. 첫 번째 단계는 최소의 지지도 이상의 발생 지지도를 가지는 조합을 찾아 빈발 단어 항목을 구성한다. 두 번째 단계는 데이터베이스로부터 연관 규칙을 생성하기 위하여 빈발 항목 집합을 사용한다. 모든 빈발 항목 집합(L)에 대해서 빈발 항목 집합의 모든 공집합이 아닌 부분집합들을 찾는다. 각각의 그러한 부분집합(A)에 대하여, 만약 support(A)에 대한 support(L)의 비율이 적어도 최소 신뢰도 이상이면, A->(L-A)의 형태의 규칙을 출력한다. 이 규칙의 지지도는 support(L)이고, 신뢰도는 support(L)/support(A)이다. Apriori 알고리즘에서 후보 집합의 생성은 Apriori-gen을 사용하여 새로운 후보 집합을 만들게 함으로써, 후보항목의 수를 줄

일 수 있다[1]. 이에 따라 연관 규칙을 찾는 시간이 감소하게 된다. R 프로그램에서는 Arules 패키지를 통해 이러한 기능을 구현할 수 있다.

3. 텍스트마이닝을 활용한 특허 검색

3.1 특허 검색 프로세스

본 논문에서 제안하는 특허 검색방법은 [Figure 2]와 같이 크게 다섯 가지의 단계로 구분할 수 있다.

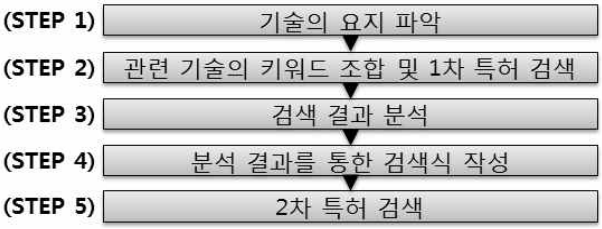
특허검색의 STEP 1에서는 조사하고자 하는 기술에 대한 이해가 필요하다. 이 단계에서는 관련 기술에 대한 이해를 통해 조사 목적과 관점을 명확히 하고, 기술에 대한 선행 조사를 통해 대표할 수 있는 키워드들을 추출하는 단계이다.

STEP 2에서는 키워드들을 조합하여 검색식을 작성하는 단계이다. 특허 검색에 있어 키워드의 수는 검색결과의 양을 결정하는 요소다. 많은 수의 키워드 조합은 사용자가 정확하게 타겟으로 하는 특허를 검색하는데 도움이 되지만, 관련기술에 대한 정확한 이해와 검색식 작성에 많은 시간이 요구되며, 관련 기술에 대한 폭 넓은 지식이 없이 작성된 검색식으로 인해 검색되지 않는 특허들이 있을 수 있다. 따라서 보통의 특허 검색에서는 검색식을 작성하고, 검색결과를 검토 후 그 결과에 따라 새로운 검색식을 작성하여 다시 검색하는 과정을 반복하게 된다. 본 논문에서는 1차 특허 검색을 통해 폭넓고 많은 수의 특허를 검색해 내는 것에 목적이 있기 때문에 몇 가지 키워드들의 조합으로만 검색식을 작성한다.

STEP 3에서는 검색 결과를 분석하는 단계이다. 분석 방법은 텍스트마이닝을 기본으로 하여 유사특허끼리의 클러스터링을 통한 후보군 도출과 그 후보군에서 키워드를 추출하고, 연관 단어 추출을 통한 복합명사 생성 과정을 거친다.

STEP 4에서는 앞 단계의 분석결과를 활용하여 검색식을 작성하는 단계이다.

마지막 STEP 5에서는 2차 특허 검색을 통해 최종 검색 결과를 도출하는 단계이다.



[Figure 2] Proposed patent search process

3.2 프로그램 구현

제안하는 특허 검색 방법은 빠른 시간 안에 사용자가 목적으로 하고 있는 특허 검색 결과 도출을 목적으로 하고 있다. 따라서 특허검색 프로세스의 STEP 2까지의 과정을 거치면 적게는 수천 개에서 많게는 수십만 개의 특허를 분석 하게 된다. 이에 따라 빅데이터 분석과 분석결과의 시각화가 용이한 R을 이용하여 프로그램을 구현하였다. 프로그램으로 구현한 부분은 제안하는 특허 검색 프로세스의 STEP 3에 포함되는 여섯 단계에 해당한다.

STEP 3.1에서 특허 문서별 corpus(글 또는 텍스트를 모아놓은 것)생성은 R의 패키지 중 하나인 tm 패키지를 활용하였다. tm 패키지는 XML, PDF 등 다양한 형식의 문서를 읽어 들여 corpus를 생성하는 기능을 제공한다.

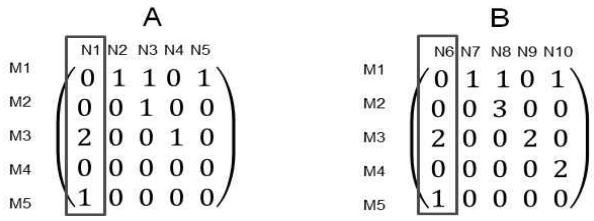
STEP 3.2에서는 생성된 문서별 corpus들에서 불용어 및 기호 등의 제거와 명사 추출을 위해 tm 패키지와 한글 처리를 지원하는 KoNLP 패키지를 사용하였다. KoNLP 패키지는 corpus 내의 한글을 형태소 단위로 인식할 수 있는 기능을 제공한다. 두 패키지의 [Figure 3]과 같은 기능을 활용하여 www, http와 같은 의미 없는 영문, 기호 등을 제거하게 된다.

```
> txt<-"시장 분석 및 컨설팅 기관인 IDC(www.idc.com)가 최근 발간한 '전세계 빅데이터(big data) 기술 및 서비스 전망 보고서'에 의하면 빅데이터 시장은 2010년 32억달러에서 오는 2015년에는 169억달러 규모에 달할 것으로 예측된다."
> extractNoun(txt)
 [1] "시장"          "분석"          "컨설팅"
 [4] "기관"          "IDC"           "www"
 [7] "idc"           "com"           "한"
[10] "'전세계"       "빅데이터(big"  "data"
[13] "기술"          "서비스"        "전망"
[16] "보고서'에"     "빅데이터"      "시장"
[19] "2010"          "년"            "32"
[22] "억"            "달러"          "2015"
[25] "년"            "169"           "억"
[28] "달러"          "규모"          "것"
[31] "예측"
> SimplePos09(txt)
$시장
[1] "시장/N"

$분석
[1] "분석/N"

$및
[1] "및/M"

$컨설팅
[1] "컨설팅/N"
```

[Figure 3] Extract noun and Hangul morpheme separated utilizing tm, KoNLP package

STEP 3.3에서는 2자리 이상의 명사만 추출하도록 프로그램을 구현하였으며, tm과 KoNLP 패키지를 이용하였다.

STEP 3.4에서는 문서의 유사도를 기반으로 한 그룹화를 smdc 패키지의 기능을 활용하여 구현하였다. smdc 패키지는 각각의 문서에서 추출한 명사들의 빈도수를 이용해 [Figure 4]와 같이 행렬로 만들어 유사도를 판단하는 기능을 제공한다.



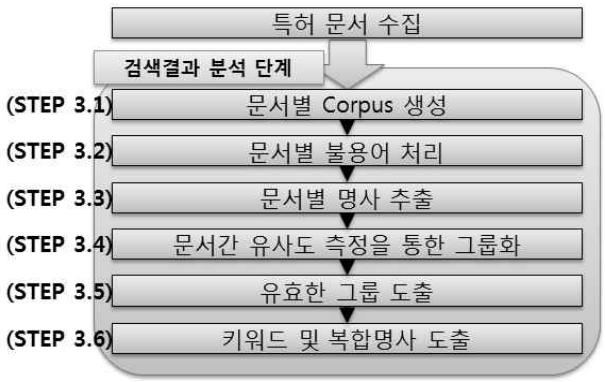
[Figure 4] comparison of the similarity matrix by the smdc package

여기서 문서 유사도 계산에 흔히 사용되는 TF-IDF 방법을 사용하지 않은 이유는 1차 특허 검색 결과에는 검색 목적에 맞지 않은 특허 문서들이 많이 섞여 있을 가능성이 있어 핵심 적인 명사들이 누락될 수 있기 때문이다.

STEP 3.5에서는 유사한 문서들로 그룹화된 특허 문서들 중에서 유효한 후보 그룹들을 선정하는 과정을 거친다. 검색의 목적에 맞지 않다고 판단되는 문서가 포함된 그룹을 삭제하는 단순한 방법을 사용하였으며, 여기서 유효한 그룹은 주관적인 판단에 의해 결정한다. 이를 통해 검색 목적에 맞는 문서들을 포함하고 있는 유효한 그룹들만 남게 된다.

마지막으로 STEP 3.6에서는 유효한 그룹에 포함된 문서들에서 키워드와 복합 명사들을 추출하게 된다. 키워드 추출을 위해서는 TF-IDF 관련 함수를 제공하는 tm 패키지를 사용하였다. 3.4단계와 달리 이 단계에서 TF-IDF를 사용하는 이유는 STEP 3.5를 통해 TF-IDF 값에 영향을 줄 수 있는 그룹들이 삭제되었기 때문이다. 복합 명사 추출을 위해서 연관성 분석 알고리즘인 Apriori의 기능을 제공하는 Arules 패키지와 이를 시각화하기 위한 ArulesViz 패키지를 사용하였다.

[Figure 5]는 프로그램의 기능 구성 및 프로세스를 보여준다.



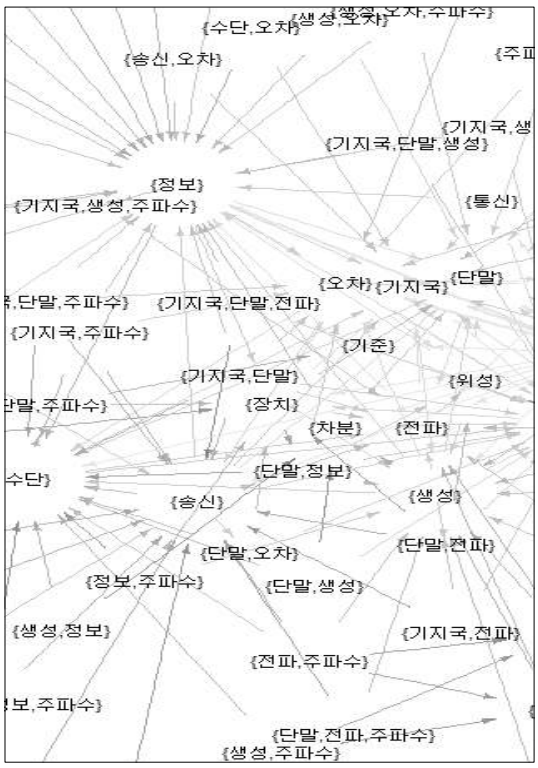
[Figure 5] Analysis process of the primary patent search results

3.3 테스트 결과 분석

테스트 환경은 쿼드코어 CPU와 8GB 메모리에 OS는 윈도우 7을 사용하는 PC를 사용하였다. R프로그램은 멀티코어 처리를 지원하는 R 3.0.2 버전을 사용하였다.

특허 검색의 목적은 위치기반 서비스 제공을 위해 사용되는 측위 방법들에 대한 국내 특허조사로 하여 1차 특허 검색을 위한 키워드로 “위치기반서비스”, “LBS”, “Location Based Service“를 사용하였으며, 특허는 WIPS를 통해 검색 하였다.

1차 검색 결과에서는 약 6만 건의 특허가 검색 되었으며, 검색 결과 중 1~300번까지의 특허의 전문을 PDF 파일로 다운로드하여 분석에 활용하였다. 300건의 특허 전문은 검색결과 분석 단계를 거치며, 15개의 그룹으로 나뉘었으며 그중 7개의 그룹은 검색 목적에 맞지 않다고 판단되어 남은 8개의 그룹에서 키워드 및 복합명사 추출을 실시하였다. 키워드 및 복합명사 추출 단계에서 복합명사의 추출은 R프로그램의 ArulesViz 패키지를 활용하였다. [Figure 6]은 유효한 8개 그룹의 문서들을 하나의 corpus로 만들어 Apriori 패키지를 통한 분석결과를 시각화한 것의 일부를 보여준다. 시각화한 결과를 통해 송신정보, 오차정보, 기지국정보, 송신오차 등 다양한 복합명사를 키워드로 활용할 수 있다.



[Figure 6] Visualization of the analysis results of the relationship

최종적으로 도출된 키워드와 복합명사를 조합한 검색식을 활용하여 WIPS로 검색하였을 때 검색목적에 맞는 약 200여개의 특허를 검색하였으며, 사용자의 수작업을 제외한 프로그램의 프로세싱 타임은 약 7분이 소요되었다.

4. 결론 및 향후 연구 과제

본 논문에서는 빅데이터 분석 도구인 R을 이용하여 빠른 시간 안에 사용자가 목적으로 하고 있는 특허검색 결과를 효율적으로 도출할 수 있는 검색어 추출에 관한 연구를 진행하였다. 이를 위해 기술의 요지 파악과 관련 기술의 키워드 조합 및 1차 특허 검색, 검색결과 분석, 분석결과를 통한 검색식 작성, 2차 특허검색과 같은 다섯 가지의 단계의 특허 검색 프로세스를 제안하였다. 제안하는 특허 검색 프로세스의 세 번째 단계인 검색 결과 분석에서는 특허 문서별로 corpus를 생성하고, 불용어 처리 및 명사추출과 문서 간 유사도 측정을 통한 그룹화 과정, 유효한 그룹만 도출하여 키워드 및 복합명사를 도출하는 과정을 프로그램화 하였다. 이러한 프로그램의 테스트 결과 검색목적에 맞는 약 200여개의 특허를 검색하였으며, 사용자의 수작업을 제외한 프로그램의 프로세싱 타임은 약 7분이 소요되어 특허의 검색에 필요한 시간을 대폭 단축시킬 수 있었다. 본 논문에서는 300개의 특허만을 가지고 분석하였으나 R프로그램에서 지원하는 다양한 빅데이터 분석 패키지를 활용할 경우 대량의 특허 분석이 가능하며, 분석 시간을 더욱 단축시킬 수 있을 것이다.

향후 연구로는 유효그룹 도출을 위한 비교 기준을 제시하여 이런 부분까지 체계적으로 분석하여 자동화 할 수 있는 방법을 구현하고자 한다.

5. References

[1] 고수정, 이정현, “Weighted Bayesian Automatic Document Categorization Based on Association Word Knowledge Base by Apriori Algorithm” 멀티미디어학회논문지 제4권 제2호, 2001. 04.
[2] 고충곤, “The Imporlance of Intellectual Property in the Coming Information Society” 정보과학회지 제17권 제10호, 1999. 10.
[3] 권동준, “특허관리전문회사(NPE)의 두 얼굴”, etnews, 2013. 05.
[4] 기획재정부, “시사경제용어사전”, 2011. 11.
[5] 김용, “A Study on Design and Implementation of Personalized Information Recommendation System based on Apriori Algorithm”, 한국비블리아학회지 제23권 제4호, 2012. 12.
[6] 윕스IP교육센터, “특허정보의 효율적 검색 및 관리”, 2011. 07.
[7] 이성직, 김한준, “Keyword Extraction from News Corpus using Modified TF-IDF”, 한국전자거래학회지 제14권 제4호, 2009. 11.
[8] 정연덕, “The Study of Patent Abuse in respect of Patent Troll”, 산업재산권 제22호, 2007. 04.

6. 저 자 소 개

장 청 윤

남서울대학교 산업경영공학과 공학사 취득. 인하대학교 산업공학과 석사 취득. 현재 인하대학교 산업공학과 박사과정 중.
관심분야 : SCM, ERP, RFID 관련 물류관리 시스템 개발 등

주소: 인천광역시 남구 용현동 253, 인하대학교 산업공학과

장 정 환

한라대학교 산업경영공학과 공학사 취득. 인하대학교 산업공학과 석사 취득. 현재 인하대학교 산업공학과 박사과정 중.
관심분야 : RFID 관련 물류 관리 시스템 개발, 항공물류 RFID 시스템 개발 등

주소: 인천광역시 남구 용현동 253, 인하대학교 산업공학과

김 석 주

현재 인하대학교 산업공학과 학부과정 중.
관심분야 : 생산관리, 공정관리, SCM 등

주소: 인천광역시 남구 용현동 253, 인하대학교 산업공학과

이 현 근

인하대학교 산업공학과 학사 취득. 한국과학기술원 산업공학과 석사, Tufts 대학 공학석사 취득. 현재 엘비세미콘 고문으로 재직 중.
관심분야 : 벤처투자, 창업활성화, 기술사업화 등

주소: 경기도 평택시 청북면 율북리 1027-1, 엘비세미콘

이 창 호

인하대학교 산업공학과 학사 취득. 한국과학기술원 산업공학과 석사, 경영과학과 공학박사 취득. 현재 인하대학교 교수로 재직 중.
관심분야 : 물류, RFID, SCM 등

주소: 인천광역시 남구 용현동 253, 인하대학교 산업공학과