결정트리 학습알고리즘을 이용한 우수특허 선별방법에 관한 연구

이수영*, 문종섭*

요약

지식재산 사회로 들어오면서 기술들의 변화 주기가 짧아 졌다. 특히 정보기술분야와 관련된 제품개발에 있어서 연구내용을 적용하여 상용화하는 시기는 매우 중요하다. 따라서 기술의 변화를 빠르게 감지하고 핵심적인 기술을 파악하기 위해서 많은 정성적 분석의 노력이 필요해지고 있다. 본 논문에서는 데이터마이닝 기술을 중심으로 빠르게 우수기술의 특허를 선별하는 방법을 제시한다. 특허 문서의 분석 요소들의 경우 계량 과 명목척도가 혼재되어 있다. 이런 점을 고려하여 특허 문서의 평가를 위해 평가요소 정리와 특징 파악, 데이터 정규화, 그리고 평가요소들 중 중요한 평가요소들을 추출하는 특징 추출을 진행하여 적절한 입력변수를 선정했다. 그리고 결정트리 학습 알고리즘을 활용하여 기술분야 전문가의 평가결과를 교사 학습하여 예측 모형을 구축했다. 그 결과 검증용 테스트데이터의 결정트리 정분류율이 70.8%로 나타났으며 트리 분석을 통해 전문가들의 우수특허선별 방법에 대해 파악해 보았다. 이를 통해서 관심 기술분야의 우수특허기술을 객관적이고 빠르게 분석할 수 있을 것이다.

A Study on Excellent Patent Selection Method using Decision Tree Learning Algorithm

Su-Young Lee*, Jong-Sub Moon**

ABSTRACT

Changes in intellectual property society, knowledge of the conversion cycle is short. Apply the product development and commercialization of research in the field of information technology, the timing is very important. Thus, changes in technology to quickly detect and identify the key technical efforts in order to more qualitative analysis, it is becoming necessary. In this paper, Excellent technique patents, how to selectively focused on data mining techniques are presented. If analysis of the elements of a patent document it is a mixture of quantitative and nominal scale. Given this, Evaluation factors for the evaluation of patent documents, filed identifying characteristics, data normalization, and a critical evaluation of feature extraction process to extract elements and appropriate input parameters were selected. And utilizing the results of expert evaluation of decision tree learning algorithm has built a predictive model of teacher learning. As a result, the results of the verification test data showed a 70.8 percent through a decision tree analysis, patent experts how to excellent screening and see if you can. Through this interest in technology the excellent patent can be analyzed objectively and quickly.

Key Words : Intellectual Property, Patent, Data Mining, Decision Tree, PCA, Cross Validation

* 고려대학교 정보경영공학전문대학원(✉leesuyoung1@naver.com)
· 제1저자(First Author) : 이수영 · 교신저자(Correspondent Author) : 문종섭
· 접수일(2012년 3월 30일), 수정일(1차 : 2012년 4월 13일), 게재확정일(2012년 4월 18일)

Ⅰ. 서 론

지식재산 사회가 도래하면서 국내 및 해외의 기술 경쟁이 치열해졌다[1,2,3,4]. 새로운 기술을 통해 경쟁에 우위를 점령하거나 새로운 기술 개발에 선두의 자리를 차지하기 위해 기업들이 기술개발 활동에 비중을 높이고 있다[5,6,7,8]. 기술개발 시작 전에 관련 연구 동향 및 기술 수준 등 관련 연구 정보 분석은 반드시 필요하다. 기술분석에 활용되는 다양한 데이터 중 특허 정보는 기업 및 연구기관의 기술개발 활동에 대한 정보를 포함하고 있는 것으로 객관적인 정보를 제공하는 소스로 인식되어 왔다[9,10]. 그러나 특허에서 추출할 수 있는 다양한 정보에 불구하고 대량의 문서에서 유용한 정보를 찾기란 쉽지 않다. 가치 있는 기술적 정보를 신속히 파악하고 효율적 관리를 위해 객관적인 분석 방법이 필요하다. 특허문서의 변화가 빠르게 이뤄지고 있으며 기술수명의 주기가 짧기에 사전 기술분석이 신속하게 이뤄져야 한다. 하지만 기술의 변화를 빠르게 감지하고 핵심적인 기술을 파악하기 위해서는 많은 정성적 분석의 노력이 필요하다. 우수특허 분석에 특허 정보의 이용이 가능한 것은 특허 자료가 장기간에 걸쳐 얻을 수 있고 규정(standard)이 객관적이고 점진적으로 변하기 때문이다[11]. 정량적 특허 분석에는 특허문서의 서지사항(bibliography) 분석, 문서내용(text mining) 분석, 특허지표 분석 등이 있다. 서지사항에 해당하는 출원일자, 등록일자, 제목, 발명자, 인용특허 등을 시계열적으로 관찰하는 등 직관적인 분석에 활용하고 있다[12,13]. 명세서의 상세한 설명, 청구범위, 해결하고자하는 문제 등의 텍스트 분석을 통해 공백기술 탐색[14], 기술예측[15] 등 텍스트마이닝 분석을 활용하여 특허문서의 내용 분석이 연구되고 있다. 그리고 특허 지표는 거시적인 측면에서 특허의 기술분석이 가능하도록 통계적인 방법을 통해 설정한 측정 방법이다[16]. 현재 대부분의 연구들이 텍스트 분석, 특허지표 개발에 중점을 두고 특허문서를 분석하여 이를 기술현황을 비교하는 연구에 이용하고 있다.

특허 문서의 분석 요소들의 경우 계량척도 와 명목척도가 혼재되어 있다. 이런 점을 고려하여 특허문서의 평가를 위해 평가요소 정리와 특징 파악이 필요하다. 특허 데이터는 각국 특허청에 데이터베이스화 되어 있다[17]. 하지만 분석 데이터로 활용하기 위해서는 데이터의 추출과 데이터 정제 그리고 데이터 정규화 과정이 필수적이다. 특허문서는 다양한 정보를 일정한 규칙으로 작성되어 있기 때문에 다양한 평가요소들을 추출할 수 있기에 이 중 우수특허 선별을 위해 중요한 평가요소들을 추출하는 특징 추출을 진행하여 적절한 평가요소를 선정한다. 그리고 우수특허 선별을 위해 결정트리 학습 알고리즘을 활용하여 기술분야 전문가의 평가결과를 교사 학습하고 예측 모형을 구축한다. 이를 통해서 우수특허기술을 객관적이고 빠르게 분석할 수 있을 것이다. 또한 특허문서의 정량적인 요소를 이용하여 객관적인 평가 방법을 제시하므로 전문가의 정성적 분석의 시간과 노력을 줄일 수 있을 것이다. 또한 거시적 차원에서 갖는 경제적 가치를 측정하고 미시적으로는 보유하고 있는 기술의 중요도 평가, 타 연구와 의 기술 중복 회피 등에 이용될 수 있다.

II. 관련연구

2.1 분류기준

본 논문에서 유사문서 추출, 평가대상 선정 등 특허문서를 분류하는 기준인 IPC(Internationl patent classification)는 특허검색, 분류, 및 정열을 위해 사용되는 국제적인 분류 기준이다. IPC는 WIPO(World Intellectual Property Organization)에서 공식적으로 관리하며 기술의 모든 분류를 다섯 단계의 구조로 분류한다. 처음 Section에서는 필수품, 수행방법, 화학,

섬유, 기계공학, 물리, 전자와 같은 8가지 구성으로 이뤄져 있다. Section은 다시 Class로 나뉜다. Class는 또 더욱 세부적으로 Subclass로 나뉜다. IPC 전부는 8개의 Section, 120 classes, 630 subclasses 그리고 69,000개의 그룹으로 이뤄져 있다[18]. 모든 특허는 특허등록을 위해 심사과정에서 전문가의 검토를 통해서 관련된 기술분야 IPC코드를 부여 받는다. 기술분야를 선정하는 개인의 영향에 따라 다른 설정이 이뤄질 수 있지만 광역적인 범위에서 IPC코드를 부여 받는다.

표. 1 IPC 분류 기준
Table. 1 IPC classfication criteria

Section	B	Performing Operations; Trans...
Class	B 64	Aircraft; Cosmonautics ...
Subclass	B 64 C	Aeroplanes; Helicpters ...
Main group	B 64 C 25/00	Alighting gear ...
sub group	B 64 C 25/02	. undercarriages
sub group	B 64 C 25/04	.. arrangement or disposition...
sub group	B 64 C 25/10	... retractable, foldable, or ...

2.2 핵심어 추출

다양한 요소가 변수로 선정이 가능하다. 청구항수, 연차등록수 등 정량적인 부분이 있으며, 유사한 특허들 간의 시간적 · 수량적 비교로서 얻을 수 있는 변수가 있다. 유사한 특허 군을 수집하기 위해 인용정보를 활용하는 것이 적합하지만 국내에서는 특허정보에 대한 인용 정보데이터의 수록의무를 강제하지 않기 때문에 인용관련 변수를 추출하기 위해서 추가적으로 자연어 처리 작업을 통해 유사한 특허목록을 구성한다.

형태소 분석 및 복합 명사 분석을 수행하여 예비 키워드를 생성했다. 예를 들어 “본 발명은 공동주택 방문자에 대한 화상 정보와 신원정보를 리스트화 하여 해당 거주자 단말기로 전송하는 네트워크 시스템을 통한 방문자 알림 서비스 시스템에 관한 것이다. “는 아래와 같이 키워드를 생성 한다.

“공동주택”, “방문자”, “공동주택방문자”, “화상”, “정보”, “화상정보”, “신원정보”, 리스트화“, ”거주자“, ”단말기“, ”거주자단말기“, ”전송“, ”네트워크(4000)“, ”시스템“, ”네트워크(4000)시스템“, ”방문자“, ”서비스“, ”시스템“, ”서비스시스템“

이어, 예비 공기쌍 정보를 추출한다. 공기쌍 정보란 위치적으로 이격되어 있지만, 함께 등장한 핵심 키워드에 대한 정보를 말한다. 상기 예시 문장에서, 다음과 같은 예비 공기쌍이 추출될 수 있다.

“공동주택방문자 | 화상정보”,“공동주택방문자 | 신원정보”,“신원정보 | 화상정보”, “리스트화 | 화상정보”, “리스트화 | 신원정보”, “거주자단말기 | 신원정보”, “거주자단말기 | 리스트화”, “리스트화 | 전송”,“거주자단말기 | 전송”,“거주자단말기 | 네트워크(4000)시스템”, “네트워크(4000)시스템 | 전송”, “방문자 | 전송”, “네트워크(4000)시스템 | 방문자”, “네트워크(4000)시스템 | 서비스시스템”, “방문자 | 서비스시스템”

핵심 키워드 추출시 상기 예비 키워드 및 예비 공기쌍에 대한 TF(Term Frequnecy)를 생성한다. 예를 들어, 거주자는 1회, 거주자단말기|네트워크시스템은 1회, 방문자는 2회의 빈도가 된다.

표. 2 핵심어 추출
Table. 2 Keywords Extraction

출원번호	WORD	제목	상세한설명	청구항	기타	TF	가중치
10-1990-0000XXX	직쇄지방산	0	0	1	8	9	8.96
	존재위치	0	0	1	3	4	8.85
	미리스트레인산	0	0	0	3	3	8.78
	팔미트레인산	0	0	0	3	3	8.78
	백색화	0	0	0	3	3	7.68
	식로	0	0	0	7	7	7.26
	파라옥시벤조...	0	0	0	3	3	7.17

예비 키워드 및 상기 예비 공기쌍들에 핵심 키워드 추출 알고리즘을 적용하여 특허 문서의 기술적 내용을 대표할 수 있는 핵심 키워드 및 핵심 공기쌍을 생성하고 저장하였다. 핵심 키워드 추출 알고리즘은 자연어 처리 기술분야에서 다양하게 제시된 방법을 사용할 수 있다. 본 논문에서는 특허문서에 해당 필드에 가중치를 부여할 수 있는 알고리즘을 사용하였다.

$$Term_{Weight} = \frac{(1+\log(1+\log(tf)) \times (1+\log(\frac{N}{df}) \times (\sum_{i=1}^{4} fw_i)}{((1-slope) \times pivot + (slope \times uf))} \quad (1)$$

tf : term frequency로 키워드(색인어)가 현재 문서에 출현한 빈도수
N : 전체 문서의 개수 , pivot : 평균 문헌 길이(ut/N)
df : document frequency로 키워드가 출현한 문서 수
slope : 기울기(임의의 상수값, 조정가능), fwi : 필드별 가중치
ut : 전체 문서집합에서의 unique terms
uf : 해당문서의 unique terms

핵심 키워드 추출 알고리즘은 하나의 특허 문서에서 추출된 용어(term)의 빈도와 그 용어를 포함하고 있는 문서의 빈도를 고려하여 추출된다[19]. 본 연구에서는 특허 문서의 발명의 명칭, 요약에 특히 가중치를 높여 특허 문서에 적합하게 중요한 키워드를 추출한다.

2.3 데이터 추출 및 전처리

수집한 다양한 데이터는 기본적인 데이터 형태 일치 작업을 진행하고 데이터 오류 보정 등 데이터 추출을 진행하여 분석 데이터로 정형화 할 필요가 있다. 기존 연구를 통해 추출된 평가요소들은 유사한 의미 또는 통계적으로 종속 변수에 미치는 영향이 유사할 수 있기 때문에 변수들 간에 관계를 분석할 필요가 있다[20].

다중공선성(Multicollinearity)은 독립변수들 간의 상관정도의 상태를 말한다. 다중공선성이 강하게 나타날 경우 데이터 분석 시 문제를 야기하는 하나의 특성으로 알려져 있다[21]. 특히 다중공선성의 클 경우 예측모형 구축에서 입력데이터의 분석에 부정적인 영향을 미치는 것으로 알려져 있다. 다중공선성은 분산팽창요인(Variance Inflation Factor, VIF), 공차(Tolerance) 등을 통해서 측정 가능하다. 평가 요소 후보군 추출 후 변수 끼리 다중공선성을 최대한 줄여서 우수특허 선별 과정에서 설명력을 높일 수 있다. 따라서 각 후보 평가요소에 대해 기초 통계량 분석과 함께 상관분석을 진행해서 상관성이 높은 평가요소를 제거하거나 모형을 재설정 하여 준다[22].

2.4 평가요소 특징추출

주성분분석(principal component)은 고차원 데이터로부터 데이터의 구조를 밝히거나, 설명력이 높은 데이터 위주로 차원을 낮추는 분석 방법이다. 변수가 많은 경우에 군집해석의 효율성을 높이기 위해 고차원의 성분을 저차원으로 변환하여 문제를 해결에 활용된다[23]. 주어진 데이터의 분산이 최대가 되는 축으로 변환하는 것으로 분산이 작은 성분을 제거함으로써 데이터의 차원을 줄이는 동시에 데이터에 포함되어 있던 잡음(noise)을 제거할 수도 있다. 또한 변수들이 보유한 정보의 손실을 최소화 하는 주성분이라 불리는 새로운 변수를 기존 변수의 선형조합으로 만들어내서 활용할 수 있다. 데이터 행렬 X의 차원을

k로 낮추는 식은 다음과 같다.

$$Y = W^k \cdot X \quad (2)$$

여기서 W는 X의 상관행렬의 고유벡터에 해당하는 고유값(eigenvalue)의 내림차순으로 정렬한 행렬이고, k는 이 중 k개의 열을 사용하겠다는 의미이다[24]. 각 고유값에 대한 고유벡터는 주성분 변수를 구성하는 결정계수를 포함하고 있다. 이를 선형결합을 통해 주성분 변수를 도출할 수 있다. 또한 주성분에 대해 가중치가 높은 변수들끼리 개체를 분류하거나 군집 분석결과에 대한 해석이나 주성분 분석에 사용이 가능하다.

2.5 분류 및 학습 방법

2.5.1 분류 및 결정트리

분류를 위한 다양한 학습 방법이 있지만 본 논문에서는 의사결정트리를 활용하려 한다. 결정트리 알고리즘은 데이터 분석의 결과가 트리구조로 표현될 수 있기 때문에 모형결과에 대한 분석이 이해하기 쉽고 학습데이터에 내포되어 있는 전문가들의 의사결정 방법을 유추할 수 있는 장점을 가지고 있다. 특허문서에서 추출한 값들뿐만 아니라 경과정보 등에서 추출한 변수들에 명목척도가 포함되어 있고 점수를 추정하는 것이 정확한 특허의 가치를 평가의 척도가 아니라 의미적인 척도가 될 것이기에 특허의 평가 등급으로 분류될 것이다[25]. 의사결정트리는 데이터마이닝의 분류 작업에 주로 사용되는 기법으로, 과거에 수집된 데이트의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 분류별 특성을 속성의 조합으로 나타내는 분류모형을 트리의 형태로 만드는 것이다. 그리고 이렇게 만들어진 분류모형은 새로운 레코드를 분류하고 해당 분류의 값을 예측하는데 사용된다. 학습을 위한 결정 트리는 종속변수를 분류하여 개별적으로 반복되는 비선형 함수를 추정하는 방법으로 이진반복 배분과정(binary recursive partitioning process)을 통해 비선형 회귀분석을 수행한다[26]. 여기서 이진반복 배분과정이란, 부모노드에서 조건에 따라 두 개의 자식노드로 이분류(binary splits)되는 과정을 트리구조에 기초하여 반복 수행하는 것을 뜻한다. 이진반복 분배 방법들 중 C4.5는 J Ross Quinlan에 의해 수정 발전된 의사결정 트리 알고리즘이다. 각 마디가 다지분리의 구조를 갖는 트리로 구성된다. C4.5는 분할자로 엔트로피지수를 이용한다. C4.5의사결정트리를 형성하기 위하여 처음 수행하는 작업이 분할정복(Divide and Conquer)이다. 입력되는 훈련 집합이 성공적으로 분할 되도록 모든 하부 집합에 하나의 클래스가 속하는 경우들로 구성될 때까지 트리를 형성한다[27].

어떤 변수를 선택하고 이 변수를 어떻게 분리하는가에 따라 다양하게 분리가 가능하다. 대표적인 분리기준은 엔트로피(Entropy), 지니(GINI), 카이스퀘어 검정(Chi-Square Statistics) 이 있다. 결정트리의 가지를 분할하기 위해서 C4.5는 정보이익비율(Information gain ratio)이 노드를 분리하는 기준으로 사용된다. 주어진 데이터를 분류하기 위하여 평균정보(Average Information)를 가장 감소시킬 수 있는 방법으로 현재의 훈련 집합을 분리한다. 이는 랜덤변수의 발생 확률이 높은 집합을 먼저 선택하는 것과 같다. 전체 훈련 집합 |S|와 함께, 현재의 훈련 집합을 S라하고, 클래스 에 속하는 경우(Case)의 수를 Freq(,S)라 하면, 주어진 예의 클래스를 확인하기 위하여 요구되는 평균정보는 식 3과 같다.

$$Info(S) = \frac{\sum_{i=1}^{n} Freq(C_i, S)}{|S|} \times \log_2\left(\frac{Freq(C_i, S)}{|S|}\right) bits \quad (3)$$

이때, 어떤 시험 X로부터 , S가 n개의 하부 집합 S1,S2..Sn으로 분리된다면, 정보이익은 식 4와 같다.

$$Gain(X) = info(S) - \sum_{i=i}^{n} \frac{|S_i|}{|S|} \times info(S_i) \quad (4)$$

여기서, 정보이익 비율은 식 5와 같이 구하게 된다.

$$Gain\,Ratio(X) = \frac{Gain(X)}{split\ \ info(X)} \quad (5)$$

순환적으로 마디가 나누어지는 과정에 따라 더 이상 개선되지 않을 때까지 트리가 형성된다. 이때 트리구조가 매우 복잡한 경우가 발생하며, 이를 데이터의 과적합(Overfitting)이라고 한다[28]. 어떤 경우에는 트리가 복잡해짐에 따라 간단한 트리보다 더 높은 에러율을 가지게 된다. C4.5의 경우 식 6와 같이 에러율을 계산한다.

$$Error\,rate = U_{cf}(E, N) \quad (6)$$

E: Event, N: Trials,
Ucf: Upper limit is the confidence for the binomial distribution,
cf : Given confidence level)

이러한 에러율을 기초로 하위트리 전체의 예측된 에러수가 상위 잎의 예측된 에러의 수 보다 크다면 상위 잎으로 대체한다. 이렇게 가지치기함으로써 예측된 에러율을 더 낮게 줄일 수 있다. 이러한 에러율은 규칙 생성에 유용하게 이용된다. 가지치기를 함으로써 크기를 최적으로 하는 의사결정트리를 형성하게 된다[29].

2.5.2 교차검증(Cross validation)

대량의 샘플데이터를 얻을 수 있다면 실제 에러율(true error rate)에 근사한 값을 측정할 수 있다. 하지만 실제 응용 단계에서는 실험에 필요한 만큼의 데이터를 확보하기 어렵기 때문에 유한한 샘플을 가지로 모델을 구축하는 것이 일반적이다. 적은 샘플데이터를 가지고 에러를 측정하기 위해서는 임의로 선택한 분류기(모델)에 전체 학습데이터를 사용하여 에러율을 추정하는 것이 가장 저렴한 방법일 것이다. 하지만 여기에는 두 가지 문제점을 가진다.

i) 모든 샘플을 가지고 학습한 최종적인 모델은 학습데이터에 최적화 되어 분류하는 과적합(Overfitting) 문제가 있을 수 있다.
ii) 추정된 에러율은 실제 에러율에 비해 낙관적으로 추축된다.

이런 문제를 해결하기 위해 더 좋은 방법은 학습데이터를 일정한 규칙을 이용해 나누어 학습하여 모델을 구축하는 방법이 있다. The holdout method, Random subsampling, K-Fold CV(Cross Validation), LOOCV(Leave-on-out Cross Validation) 등의 다양한 방법이 있다[30]. 본 논문에서는 가장 일반적으로 사용하는 K-Fold 교차검증 방법을 활용하는 연구방법을 사용하고자한다. K-Fold 방법은 전체 샘플데이터를 K부분으로 나누어 그림1 과 같이 각 실험마다 학습데이터와 테스트 데이터로 구분하여 학습과-검증과정을 거친다. 다시 말해서 분할된 부분집합 중 한 개의 부분집합은 독립적인 시험자료로 사용하고, 나머지 K-1개의 부분집합을 모형구성 과정에서 이용되는 학습자료로 사용한다. 이와 같은 과정은 모든 모형구성 과정에서 총 K번 반복되며 각 경우 에서 각기 다른 부분집합이 학습데이터로 선택된다.

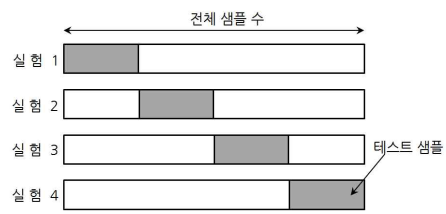

그림 1. K-Fold
Fig. 1 K-Fold

교차확인 분석이 실시되면 최종적으로 N개의 다른 모형이 구축되며, 이들 N개의 정확도를 평균한 값이 본 모형의 가장 좋은 정확도가 된다.

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i \qquad (7)$$

교차검증는 학습모형을 구축하는 각 단계 마다 에러율을 파악하므로 다양하게 활용할 수 있다. 본 논문에서는 교차검증은 모형구축 시 모형 선택, 알고리즘, 모형 파라미터 선정 그리고 적절한 변수선택에까지 다양하게 활용이 가능하다. 본 논문에서는 모형 선별 과정에 알고리즘 및 파라미터 선정 그리고 학습데이터 검정에 이용하였다.

III. 연구방법

3.1 제안방법

본 연구에서는 우수특허를 선별하기 위해 그림 2와 같이 평가 방법을 구축하였다. 각 해당하는 특허의 변수에 해당하는 데이터 추출을 특허문서의 서지사항 및 경과정보 등을 통해서 추출하였다.

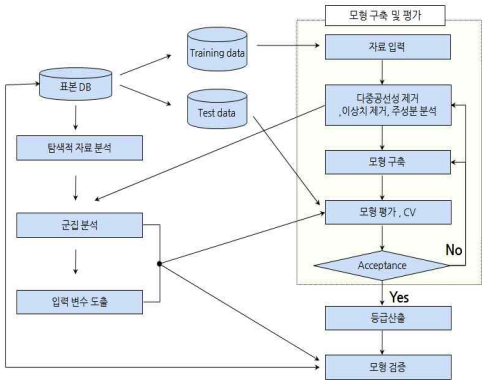

그림 2. 제안방법
Fig. 2 The proposed method

그리고 변수에 통합할 자연어처리를 통해 핵심어 추출을 연구방법대로 추출하고 데이터의 유형 정리와 기초통계 분석을 수행 하여 모형 구축에 필요한 학습데이터를 구축하였다.

3.2 변수설정

3.2.1 목표 변수

학습데이터를 통해 평가모형을 구축하기 위해 전문가들의 평가를 각 기술분야별로 수행하여 표3과 같이 수집하였다. 권리, 기술, 시장성의 전문가가 특허에 대해 점수를 기술하고 총점을 합산하는 방법으로 설문을 진행하였다. 전문가 평가로 수집된 점수는 평균이 52.7이고 편차가 7.65인 정규분포를 이루고 있다. 이를 정규분포 기준으로 표4와 같이 3개 등급(A,B,C)으로 나누어 목표변수로 설정하였다.

표. 3 기술분야별 조사 건수
Table. 3 Number of technology field survey

기계	물리	바이오	IT	화학	기타	계
150	150	150	250	150	150	1,000

표. 4 전문가평가 등급별 구간
Table. 4 Classes of expert evaluation rating

구분	A등급	B등급	C등급	계
등급 구간	0~45	46~62	63~100	1,000
개 수	92	734	174	

3.2.2 입력 변수

분석 관점에 따라 다양한 입력변수를 추출할 수 있다. 특히 구조화된 특허문서에는 정량화 할 수 있는 다양한 입력변수가 존재한다[18,20]. 본 논문에서는 입력변수 후보로 아래 표와 같이 추출 했다.

표. 5 입력 변수
Table. 5 Input variables

변수명	변수 설명
청구항수	서지정보의 청구 항 수
국내 패밀리특허	분할출원이나 이중출원/변경출원 수
권리자유사특허점유율	유사특허군내에서 평가대상 특허권자의 점유율
총피인용수	배경기술/선행기술/의견제출통지시의 피인용수
기술의 경제적수명	IPC subclass 레벨에서의 기술분야별 진부화 계수를 사용하여,특허의 실질적 잔여 수명 계산
수명주기상의 위치	전체 핵심 기술 키워드의 시간적 분포를 고려할 때 대상 특허에 포함된 핵심기술 키워드들의 상대적인 시간 위치값을 계산함
연차등록수	해당 특허의 연차등록료 납입 횟수
실시권자수	유효한 등록 실시권자 수
당사자계 심판	무효심판과 권리범위(적극적, 소극적) 심판
관련특허등록 증감의 시계열적 패턴	유사특허 년도별 증감을 시계열적으로 파악
:	:

3.3 평가요소의 유의성 검증

평가 모형 구축 시 사용되는 평가요소 후보군인 독립변수들은 기존 특허지표연구 및 전문가의 의견을 통해 추출하였다. 다양한 개념을 통해 평가요소 후보군을 추출하였기에 의미적으로 높은 연관성을 지고 있을 수 있다. 독립변수 간에 이런 상관성은 우수특허 선별을 위한 모형의 성능에 영향을 미친다. 따라서 변수간의 관련성을 진단하는데 널리 사용되고 있는 VIF(분산팽창계수)를 사용하여 다중공선성을 측정 하였다.

실험 결과 표6과 같이 분산팽창계수의 수치가 높이 나타내는 변수 ID032, ID033 대해 의미를 파악한 결과 관련 특허권자의 성격비율에 대한 의미를 내포하고 있었다. 이에 두 변수 중 대표변수로 ID032를 선택하고 ID033번을 제외하여 평가 모형 구축을 진행했다.

표. 6 변수간 분산팽창계수
Table. 6 Variance Inflation Factor

ID	ID001	ID002	...	ID032	ID033	...
VIF	2.440	1.596	...	15.767	13.890	...

추가적으로 평가요소 후보군의 우수특허 선별을 위한 가장 큰 설명력 확인하고 이상치 제거 및 추후 평가 모형 구축의 성능을 향상과 변수들 간의 분류를 파악하기 위해 주성분 분석을 수행하였다. 주성분 분석은 공분산 행렬을 이용하여 분석하지만 독립변수의 기초 통계량을 분석한 결과 각 평가요소 후보들 간의 측정 단위가 상이하게 나타났다. 따라서 주성분 분석 시 단위를 정규화 하기 위해 상관행렬을 이용하였다. 앞서 언급한 다중공성선 문제를 제거한 독립변수 41개의 주성분을 분석한 결과 고유치가 1을 넘는 대리변수가 22개로 나타났다. 가장 큰 고유치를 갖는 주성분의 분산 설명력은 10.2%이며, 그림 3에 보이는 대로 고유치 1을 넘는 대리변수들의 설명력이 71.4% 보이고 있다.

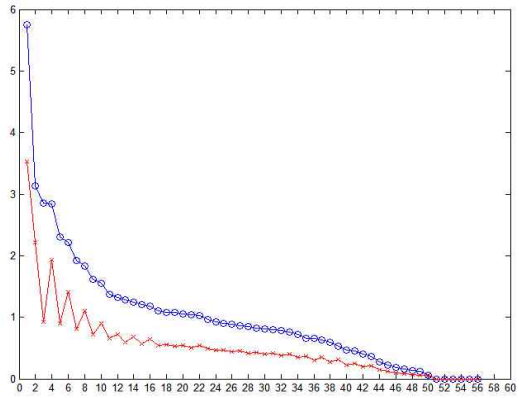

그림 3. 스크리 도표
Fig. 3 scree plot

정확률과 MAE/RMSE 값 그리고 결정트리의 구조 등을 복합적으로 고려해서 교차검증 단계에서 적은 최적의 성능을 보이는 모형을 구축할 수 있는 가장

높은 설명력을 보여주는 주성분 변수 11개를 선정하였으며 각 주성분 변수들의 고유벡터를 분석하여 표 7과 같이 그룹화 및 그룹의 성분을 검토 하였다.

표. 7 주요 주성분 변수
Table. 7 The main principal components variables

주성분	그룹이름	설명력(%)
1	게시기술의권리화정도	10.29
2	권리확인심판수	6.12
3	심사청구경과일수	5.70
4	O/A회수	5.29
5	출원인형태	4.76
6	독립청구항	4.10
7	관련분야시장의기여도	3.71
8	관련분야시장의매출예측치	3.31
9	명세서정량수치	2.87
10	O/A관심도	2.74
11	총피인용수	2.67

IV. 실험결과

4.1 모형구축

우수특허 선별의 예측모형을 모델링하기 위해 앞서 설명한 바와 같이 수집된 학습자료를 통해 학습한 결과 SMO, Bagging과 비교해서 결정트리 알고리즘인 C4.5 알고리즘을 이용한 모형 모델링을 위한 학습 데이터의 분류율은 83%로 나타나며, 주성분을 이용한 데이터의 경우 80%로 훈련데이터에 대해 정확하게 분류하고 있다. 주성분 변수를 활용했을 때 적은 변수로 유사한 성능을 보이는 분류 모형을 제시할 수 있음을 알 수 있다.

표. 8 모형구축 결과
Table. 8 Result of modeling

예측 방법	SMO	Bagging	C4.5	PCA→ C4.5(11)
정분류	73.73%	73.73%	83.07%	80.93%
오분류	26.27%	26.27%	16.93%	19.07%
Kappa statistic	0.0355	0.0355	0.515	0.4281
MAE	0.291	0.291	0.1822	0.1899
RMSE	0.3775	0.3775	0.3018	0.3081

예측모형의 세부적인 평가결과는 ROC-커브 그래프에서 보이는 것처럼 C4.5 일반모형의 경우 A의 등급을, 11개의 주성분변수를 이용한 경우 C등급을 더 잘 분류하고 있음을 알 수 있다.

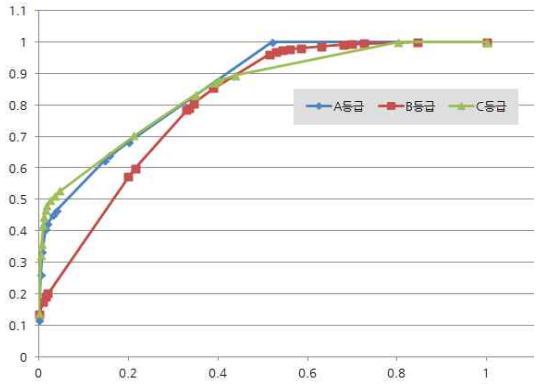

그림 4. C4.5 모형 ROC 도표
Fig. 4 ROC plot of C4.5 Model

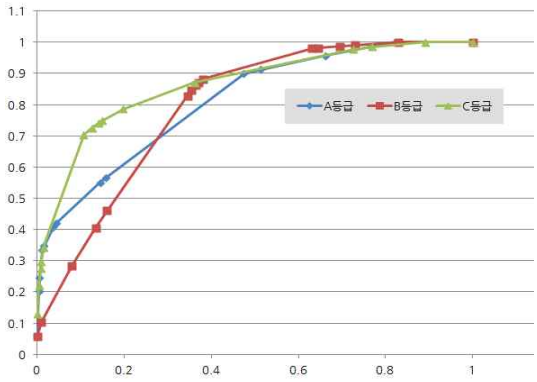

그림 5. PCA 이용 ROC 도표
Fig. 5 ROC plot of Model using PCA

결정트리 구성은 관련연구 2.5절의 분석에서 논의된 방법에 따라 진행되었다. 과도적합의 위험을 피하기 위해 가지치기를 실시하였으며 기존의 독립변수의 트리는 종료규칙에 의해 42개의 잎(leaf), 76개의 트리를 가진 결정트리가 생성되었다. 결정트리 분석결과 등급이 'A' 이상으로 판단하는 노드를 분석한 결과를 살펴보면 다음과 같다.

i) 공동출원인 많고(>0), 청구항 수가 적고(<=7), 청구항 수가 크고(>2), 심사청구 경과일 크고(> 90)
ii) 공동출원인 많고(>0), 청구항 수가 많고(>7), 해외 패밀리 특허수가 적고(<=2), (진부화)기술잔여수명(<=1.6)
iii) 공동출원인 많고(>0), 청구항 수가 많고(>7), 해외 패밀리 특허수가 많고(>2) , 연차등록회수가 적은(=1), 관련 업종의 시장크기(<=45,649,966)
v) 공동출원인 많고(>0), 청구항 수가 많고(>7), 유사특허군에서의 피인용 백분위 비율(>75, <=84)

이를 통해 다음과 같은 해석이 가능하다.

1) 공동출원인이 많을 경우 관련 연구자가 많아짐으로 우수 특허로 판단
2) 청구항 수가 많을 경우 기술적인 면에서 우위를 가질 수 있다고 판단
3) 해외 패밀리 특허수가 많으며 특허의 연차등록회수가 적고 관련업종의 시장크기가 비교적 적을 경우 시장의 초기단계에 최신의 우수 특허로 판단
4) 최신의 특허와의 키워드 일치성이 적고, 기술잔여 수명이 적은 특허일 경우 우수 특허라고 판단
5) 청구항수가 많으며 유사특허군에서의 선행기술이 적을수록 우사한 특허로 판단

4.2 모형 검증

본 논문에서는 알고리즘 선택, 파라미터 선별 그리고 학습데이터의 교차검증을 통해 모형구축 모델링을 최적화 하였다. 750건의 샘플데이터를 학습데이터로 활용했으며, 250건은 학습검증에 활용하였다. 학습데이터에 대한 분류결과는 83%의 정분류율의 결과 값을 보여 주었다. 그리고 10-Fold 교차검증 결과 68.4%의 정분류 결과를 나타내었으며 실제 테스트데이터의 결과에서는 70.8%로 최종검증에 유의미한 결과를 보여 주었다. 알고리즘 선택과, 파라미터 선별 각 단계에 교차검증을 적용하여 나온 결과값을 기반으로 최적의 모형구축을 수행하였다. 또한 실제 학습에서 보인 주성분의 변수를 활용한 모형구축은 실제 정보를 기반으로 구축한 결과와 유사한 예측율을 보여주고 있으므로 적은 변수를 통해서 같은 성능의 모형을 구축할 수 있음을 보여주고 있다.

표. 9 모형검증 결과
Table. 9 The results of model verification

예측 방법	학습데이터(750)		10-Fold CV		테스트데이터(250)	
	C4.5	PCA->C4.5	C4.5	PCA->C4.5	C4.5	PCA->C4.5
정분류	83.07%	80.93%	68.40%	70.80%	70.80%	66.80%
오분류	16.93%	19.07%	31.60%	29.20%	29.20%	33.20%
Kappa statistic	0.515	0.4281	0.1671	0.1857	0.1478	0.0391
MAE	0.1822	0.1899	0.2531	0.2531	0.2564	0.2701
RMSE	0.3018	0.3081	0.4137	0.3984	0.4043	0.419

V. 결론 및 향후 연구 방향

최근 특허의 중요성 및 특허 자료에 대한 접근성이 용이해지고 있으며 특허분석의 실용성이 높이 평가 받고 있다. 이에 본 연구에서는 특허문서를 분석하여 우수한 특허를 선별할 수 있는 분류 방법을 제시했다. 특허문서의 평가요소로 활용할 수 있는 기존에 연구된 변수들을 선별하고 이에 대한 분석과 의미를 파악하여 활용했다. 특히 평가요소의 주성분분석과 교차검증을 통해 최적의 정확률을 보이는 새로운 주성분 변수 11개에 대한 구조와 의미를 파악할 수 있었다. 그리고 유사문서를 추출하기 위해 특허문서의 핵심어 추출 방법을 제시하고 주성분분석을 통해 다양한 변수들을 군집화 하여 그에 따른 의미를 파악하였다. 독립변수들의 기초 통계량과 전문가들이 우수특허를 선별하는 성향을 파악하기 위해 분류 및 학습 방법으로는 비모수적인 방법의 결정트리 학습알고리즘을 활용하였다. 다양한 독립변수에서 평가 시 주의 깊게 파악하는 결정적인 변수를 파악할 수 있었다.

본 연구에서는 입력변수인 평가요소의 전처리를 통해 다중공선성 그리고 주성분 분석을 활용해서는 변수들의 구조관계를 파악하는 방안에 대한 방안을 제시하였다. 또한 평가모형인 결정트리 학습알고리즘의 과적합을 교차검증을 통해 분석하여 최적의 평가모형을 구축하였다. 결정트리 학습알고리즘을 적용한 모형은 다른 모형에 비해 분석이 용이하며 트리구조로 결과가 나오기 때문에 의사결정에 직접적으로 활용이 가능하다. 특히 본 연구에서 활용한 결정트리 모형은 지식재산권에 대한 경영자의 의사결정과 특허의 유지 및 포기에 직접적으로 활용이 가능하다는 장점을 가지고 있다. 따라서 추후 연구 방향으로는 다양한 평가모형을 활용하여 모형을 구축하여 결정트리 모형의 단점인 과적합을 해결해야 할 것이며 또한 독립변수가 예측결과에 미치는 요인을 분석하여 모형에 반영할 필요가 있을 것이다. 모형구축과 함께 평가요소로 활용할 수 있는 독립변수의 연구가 지속적으로 진행되면 더 정확한 예측모형을 구축할 수 있을 거라 생각한다.

참고문헌

[1] 이기식, "A Study on Institutional Aspects of Korea's Copyright Policy through USA's Experience in the Digital Age", 지방정부연구, 6(1): 257-275, 2002.

[2] Oong-Hyun Sung, Kyeong-Seon Jo, "A Study about the Effects of Intellectual Property Investment and Management on the Value of Intangible Assets of Firms", *Journal of Korea Technology Innovation Society*, 12(2): pp291-311, 2009.

[3] Yoon Byung-seop, "Infringement Status of Overseas Intellectual Property Right and Required Strategy", *Journal of Korea Technology Innovation Society*, 11(1): 23-45, 2008.

[4] 제대식, 이은철, 윤국섭, *지식경영과 특허전략*, 세종서적, 2000

[5]"중소기업 연구인력 지식재산 교육 열기 확산 등.", *Journal of the KSME*, 51(3): 53-55, 2011.

[6] Dong-Woo Yang, Da-Jin Kim, "Causal Relationship between Firms; R&D Collaboration and Performance in Contents Industry", *JOURNAL OF THE KOREA CONTENTS ASSOCIATION*, 10(4): 306-316, 2010.

[7] 이명규, "지식재산 보호는 新국가경쟁력.", *Intellectual Property right*, 46-48, 2009.

[8] 이우성, *R&D 투자의 경제적 파급효과 분석*, KISTEP 세미나 발표자료, 2008.

[9] Hall, B.H., A. Jaffe, and M. Trajtenberg, "Mar-ket Value and Patent Citations, *The RAND Journal of Economics*, Vol.36, No.1, pp 16-38, 2005.

[10] Shin Seunghoo, Hyun Byunghwan, "A Study of Analysis Methods on R&D Productivity Using Patent and Paper Analysis", *Journal of Korea Technology Innovation Society*, 11(3): 400-429, 2008.

[11] 권도윤, 이응봉, "A Study on the Design of Metadata for Patent Information Management System", Journal of the Korean Society for Information Management, 377-387, 2003.

[12] Dong-Min Kim, Yoon-Ji Choi, Chil-Woo Lee, "Analysis the Mobile User-Interface in Patent", *JOURNAL OF THE KOREA CONTENTS ASSOCIATION*, 11(12): 455-465, 2011.

[13] Do-Hoe Kim, Sang-Sung Park, Young-Geun Shin, Dong-Sik Jang, "Patent Analysis of Information Security Technology for Network-Centric Warfare", *JOURNAL OF THE KOREA CONTENTS ASSOCIATION*, 7(12): 355-364, 2007.

[14] Sunghae Jun, Sang-Sung Park, Young-Geun Shin, Dong-Sik Jang, HoSeok Chung, "Forecasting Vacant Technology of Patent Analysis System using Self Organizing Map and Matrix Analysis", *JOURNAL OF THE KOREA CONTENTS ASSOCIATION*, 10(2): 462-480, 2010.

[15] Jinho Choi, Heesu Kim, Namgyu Im, "Keyword Network Analysis for Technology Forecasting", *Journal of Intelligent Information Systems*, 17(4): 227-240, 2011.

[16] Han-Seop Shin, "Design of Consolidated Patent Index for Effective Utilization of Patent Information", *Korean Management Science Review Special*, 24(2): 1-18, 2007.

[17] 유재복, *특허정보조사의 이론과 실제*, 형설서적, 2004.

[18] Heon Kim, Dong-hyun Baek, Min-ju Shin, Dong-seok Han, "A Model for Evaluating Technology Importance of Patents under Incomplete Citation", *Journal of Intelligent Information Systems*, 14(2):121-136, 2008.

[19] Ji Seoung Kim, Joon Ho Lee, Sang Ho Lee, "Performance Evaluation and Analysis of Recent Information Retrieval Models", *Journal of KISS : Databases*, 28(2):266-278, 2001.

[20] Jiyoun Lim, Chulyoung Kim, Jachul Gu, "Analysis of Causal Relationship between Patent Indicators and Firm Performance", *Korean Management Science Review Special*, 28(2): 63-74, 2011.

[22] Donald E. Farrar and Robert R. Glauber. "Multicollinearity in regression analysis: The problem revisited", *The Review of Economics and Statistics*, 49(1):92-107, 1967.

[22] Su-dong Park, Oong-hyun Sung, "Estimation of S&T Knowledge Production Function Using Principal Component Regression Model", 13(2):231-251, 2010.

[23] 권세혁, *다변량 데이터 분석과 활용*, 자유아카데미, 2008.

[24] I.T Jolliffe(2002), *Principal Component Analysis*, Springer, 2002.

[25] Jae-Bok Yoo, Young-Mee Chung, "Analysis of Factors Influencing Patent Citations", *Journal of the Korean Society for Information Management*, 27(1): 103-118, 2010.

[26] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., *Classification and regression trees The Wadsworth Statistics/Probability Series,* Wadsworth International group, California, USA, 1984.

[27] Quinlan J.R, *C4.5: Programs for Machine Learning,* Morgan Kaufmann, 1993

[28] Levin, N. and J. Zahavi, "Predictive Modeling Using Segmentation", *Journal of Interactive market-ing*, Vol.15, pp. 2-22, 2001.

[29] Minsoo Lee, Young Chan Choe, Byungjoon Yoo, "Identifying Early Adopters of Information Systems by Inductive Learning Using Decision Tree Method", *Information Systems Review*, 9(1):67-84, 2007.

[30] Andrew W. Moore, "Cross-validation for detecting and preventing overfitting"

저자소개

이수영(Su-Young Lee)

2006년 고려대학교 전자 및 정보공학 (공학학사)

2008년~현재 한국발명진흥회 연구원

※ 관심분야 : 데이터분석, 특허, 데이터마아닝

문종섭(Jong-Sub Moon)

1981년~1985년 : 금성 통신 연구소 연구원

1991년 : Illinois Institute of technology 졸업(전산학 박사)

1993년~현재 : 고려대학교 전자 및 정보공학부 교수

※ 관심분야 : 생체인식, 침입탐지, 운영체제