한국지능시스템학회 논문지 2011, Vol. 21, No. 3, pp. 341-346
DOI : 10.5391/JKIIS.2011.21.3.341

앙상블모형을 이용한 공백기술예측

Vacant Technology Forecasting using Ensemble Model

전성해

Sunghae Jun

청주대학교 바이오정보통계학과

요 약

공백기술예측은 기술경영 분야에서 중요하게 다루어지는 주제이다. 다양한 분야에서 현재까지의 기술개발결과를 분석하여 상대적으로 연구개발이 이루어지지 못한 분야를 찾아내어 개발하는 것은 국가와 기업의 발전에 중요한 영향을 미친다. 현재 특허는 기술개발결과에 대한 가장 객관적인 데이터 중 하나이다. 본 논문에서는 특허데이터를 이용하여 공백기술을 정량적으로 예측할 수 있는 방법에 대하여 연구한다. 하나의 정량적 기술예측모형이 완벽하다는 보장을 할 수 없기 때문에 본 연구에서는 여러 가지 모형들의 결과를 결합하여 예측하는 앙상블모형을 제안한다. 통계적 분석기법과 기계학습 알고리즘을 결합하여 보다 객관적이고 정확한 공백기술예측모형을 구축한다. 제안방법의 객관적인 성능평가를 위하여 각 기술분야에 대하여 최초 특허가 이루어진 시점부터 최근까지 출원, 등록된 특허데이터를 이용한다.

Abstract

A vacant technology forecasting is an important issue in management of technology. The forecast of vacant technology leads to the growth of nation and company. So, we need the results of technology developments until now to predict the vacant technology. Patent is an objective thing of the results in research and development of technology. We study a predictive method for forecasting the vacant technology quantitatively using patent data in this paper. We propose an ensemble model that is to vote some clustering criteria because we can't guarantee a model is optimal. Therefore, an objective and accurate forecasting model of vacant technology is researched in our paper. This model combines statistical analysis methods with machine learning algorithms. To verify our performance evaluation objectively, we make experiments using patent documents of diverse technology fields.

Key Words : 공백기술예측, 특허데이터, 통계적분석기법, 기계학습알고리즘, 앙상블모형

1. 서 론

기술예측(technology forecasting)은 분야에 따른 향후 필요기술에 대한 개념을 이끌어내는 작업이다. 기술예측은 델파이(Delphi)를 통한 전문가집단의 활용 등 다양한 관점에서 접근할 수 있지만 본 논문에서는 특허데이터의 분석을 통하여 객관적인 정량적 분석에 바탕을 둔 접근방법을 제안한다.

과학기술기본법(일부개정 2010. 2. 4. 법률 9992호) 제13조(과학기술예측 등) 제1항은 다음과 같다[1]. "정부는 주기적으로 주요 과학기술통계와 지표를 조사·분석하고 과학기술의 발전 추세를 예측하여 그 결과를 과학기술정책에 반영하여야 한다." 이를 통하여 정부는 국가연구개발사업의 관리를 위하여 기술예측이 반드시 필요함을 법적으로 명시하고 있다. 이와 같은 과학기술예측조사, 미래유망기술의 발굴 등은 일본, 유럽, 미국 등 전 세계의 대부분의 국가에서 정책적으로 중요하게 다루어지고 있다[2-4]. 현재 기술예측을 위하여 델파이와 같은 전문가집단의 판단에 의한 주관적인 방법이 사용되고 있지만 다른 한편으로는 객관적인 접근방법도 시도되고 있다[5-9]. 지금까지 가장 활발하게 연구, 활용되고 있는 객관적 예측방법은 특허지도(patent map)를 이용하는 것이다[10-11]. 특허지도는 "출원된 특허정보를 가공하여, 그 분석결과를 그래프 등으로 시각화 한 것"으로 정의된다[12]. Micropatent사의 Aureka, Thomson Corporation사의 Thomson Data Analyzer 등 전 세계적으로 특허지도와 같은 특허분석결과를 제공해 주는 다양한 분석도구들(tools)이 있다. 국내에도 특허청이 개발하여 보급한 특허분석도구인 PIAS(patent information analysis system)가 있다[12]. 현재 사용되고 있는 대부분의 특허분석도구는 특허문서에 포함된 데이터의 요약(표 등)과 단순한 시각화(그래프 등)에 의존하고 있다. 그러므로 기술예측이라는 정교하고 복잡한 작업을 수행하기에는 한계가 있다. 이와 같은 문제를 해결하기 위하여 다변량통계분석(multivarate data analysis), 자기조직화지도(self organizing map), 매트릭스분석(matrix analysis) 등 보다 정교한 모형을 이용한 특허데이터의 분석에 관한 연구결과들이 나타나고 있다[7,13-14]. 이와 같은 최근의 연구결과들은 모두 한 개의 우수한 모형을 이용하여 기술예측을 시도하였다. 가장 적합하다고 판단된 한 개의 모형을 선택하여 기술예측을 수행하는 것도 의미 있는 작업이지만, 본 연구에서는 서로 다른 몇 개의 모형을 결합한 기술예측을 수행을 통하여 보다 객관적이고 일반화된 예측결과를 얻기 위하여 노력하였다. 이를 위하여 본 논문에서는 우수한 여러 개의 모형결과를 결합하는 앙상블 모형을 제안한다. 즉, 기술예측에 적합한 모형들을 찾아내어 이들의 효과적인 결합방안을 제시한다. 특히 기술특허들의 분류를 위한 군집화

접수일자 : 2011년 3월 16일
완료일자 : 2011년 6월 1일

(clustering)를 위하여 통계학과 기계학습 분야의 여러 기법들을 결합하였다. 본 논문에서 제안하는 방법의 성능평가를 위하여 미국특허청으로부터 다양한 기술분야의 특허문서 데이터를 이용한다. 최초 등록된 특허부터 일정기간동안의 특허문서를 학습데이터(training data)로 사용하였고 이후부터 현재까지의 특허문서를 테스트 데이터(test data)로 사용하여 제안방법의 성능평가를 수행하였다.

2. 공백기술예측

공백기술(vacant technology)은 하나의 기술영역에서 해당기술을 세부 기술분야로 분류했을 때 다른 분야에 비해 상대적으로 연구와 개발이 제대로 이루어지고 있지 못한 세부기술로 정의된다[14]. 최근에 이와 같은 공백기술예측에 대한 필요성에 의해 해당기술을 출원, 등록한 특허문서를 분석하여 새로운 기술경향을 찾으려는 시도가 이루어지고 있다[7,14]. 아울러 주관적인 예측보다 객관적인 특허데이터의 분석을 통하여 기술을 예측하려는 연구결과들도 나타나고 있다[9,15]. 현재 사용되고 있는 공백기술 예측의 방법들은 다음과 같이 여러 가지 관점에서 접근할 수 있다. 예를 들어, 하나의 기술에 대한 세부 기술분류를 통하여 다음 그림과 같은 매트릭스 형태의 결과를 얻을 수 있다[7].

	A1	A2	A3
B1	110	121	94
B2	57	35	117
B3	216	63	7

그림 1. 매트릭스에 의한 세부기술 분류
Fig. 1. Detailed technology classification by matrix

위의 결과는 하나의 기술분야를 크게 2개의 범주(A, B)로 나누고, 다시 각각의 기술 A와 B를 각각 3가지의 세부기술(A1, A2, A3 / B1, B2, B3)로 나눈 것이다. 해당 세부기술 코드가 만나는 곳(cell)의 수는 기술개발 건수를 나타낸다. 일반적으로 객관적인 기술개발건수로 특허건수를 사용한다. (A1 and B3)에 해당되는 세부기술의 개발건수가 216건인데 비해 (A3 and B3)에 해당되는 세부기술의 개발건수는 7건에 불과하다. 때문에 (A3 and B3)에 해당되는 세부기술은 현재시점에서의 공백기술로 판정될 수 있다. 또 다른 공백기술은 다음과 같은 주성분 산점도(principal component plot)에 의해 찾을 수 있다[14].

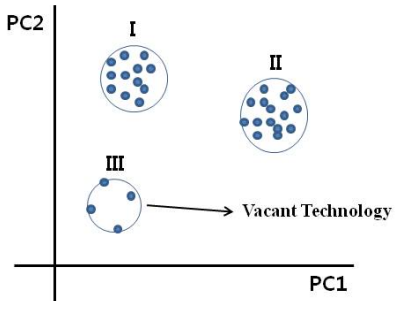

그림 2. 공백기술예측을 위한 주성분산점도
Fig. 2. Principal component plot for vacant technology forecasting

위 그림은 모든 특허를 제1주성분(PC1)과 제2주성분(PC2) 점수로 나타내고 이를 2차원 산점도로 나타낸 것이다. 군집 I과 II에 비해 군집 III은 4개의 기술특허만을 포함하고 있다. 따라서 군집 III에 속한 4개의 기술을 통하여 해당 기술분야의 공백기술을 정의할 수 있다. 그림1에 비해 그림2에 의한 공백기술예측은 공백기술의 정의를 위하여 전문가의 주관적인 판단이 추가적으로 요구된다. 즉 군집 III에 속한 4개의 특허를 대표할 수 있는 기술을 정의해야 한다.

본 논문에서는 이와 같은 공백기술예측의 성능을 향상시키기 위하여 좀 더 객관적인 특허분석 기법을 제안한다. 객관적인 기법으로서 제안되는 연구에서는 한 개의 예측모형이 아닌 다수의 예측모형을 생성, 선택, 결합하는 앙상블(ensemble) 모형을 통하여 공백기술의 예측성능을 향상시키고자 하였다.

3. 공백기술 예측을 위한 앙상블모형

기술예측은 일반적인 예측모형에 비해 좀 더 복잡하고 어려운 작업이다. 기술예측에서 사용되는 대표적인 데이터인 특허문서는 일반적인 예측모형에서 사용되는 양적(quantitative) 또는 질적(qualitative)데이터가 아니라 문서(document) 형태로 존재한다. 뿐만 아니라 기술에 대한 예측은 매우 다양한 요인들이 서로 복잡하게 작용되기 때문에 특정한 한 개의 모형을 통해 예측한다는 것은 매우 어렵다[16]. 이와 같은 어려움을 극복하기 위하여 본 논문에서는 우수한 다수의 모형들로부터 분석결과를 결합(voting)한 앙상블모형을 구축하여 예측을 수행한다. 우선 텍스트 마이닝의 전처리(preprocessing)과정[17]을 통하여 다음 그림의 첫 번째와 같이 단어행렬(term matrix)을 만든다. 이 행렬은 각 특허문서의 초록(abstract)으로부터 구할 수 있다. 단어행렬의 행은 개개의 특허문서를 나타내고 열은 n개의 특허문서에 나타난 단어를 표시한다. 즉 2번째 특허문서(Doc 2)에는 첫 번째 단어(Term 1)가 1번 사용되었음을 알 수 있다.

Term Matrix

	Term 1	Term 2	…	Term k
Doc 1	0	2	…	0
Doc 2	1	0	…	0
…	…	…	…	…
Doc n	0	0	…	1

(k >> p)

SVD – Principal Component Matrix

	PC 1	PC 2	…	PC p
Doc 1	-2.30	17.74	…	0.18
Doc 2	4.26	0.39	…	-1.57
…	…	…	…	…
Doc n	-10.89	-4.89	…	15.38

그림 3. 단어행렬과 주성분행렬
Fig. 3. Term matrix and principal component matrix

위 그림의 단어행렬에서 대부분의 행과 열이 만나는 셀(cell)의 값은 0이다. 이 문제점 때문에 어떠한 예측모형도 바로 적용하기에는 어려움이 있다. 본 연구에서는 SVD-주성분행렬(SVD-principal component matrix)을 이용하여 이 문제를 해결한다. 일반적인 주성분분석이 아니라 비정칙분해(singular value decomposition; SVD)[18]에 의한 주성분분석을 사용하는 이유는 일반적인 특허문서 데이터로부터 얻게 되는 단어행렬의 특성 때문이다. 단어행렬에서 특허문서의 수 n보다 단어의 수 k가 일반적으로 훨씬 크다. 본 연구의 실험에서 사용될 실제 특허데이터로부터 구한 단어행렬에서도 특허문서의 수(67)보다 단어의 수(1349)가 훨씬 크게 나타났다. 단어행렬에서 단어와 문서는 각각 주성분분석에서 변수(variable)와 관측치(observation)에 해당된다. 일반적인 주성분분석에서는 관측치의 개수가 변수의 개수에 비해 커야만(n>>k) 주성분분석이 가능하다[19-21]. 이 조건을 만족해야 변수의 개수만큼 주성분을 구해주고 이 결과를 통하여 설명력이 가장 큰 제1주성분부터 몇 개의 주성분을 사용하여 추가적인 분석모형을 구하게 된다. 그러므로 특허데이터로부터의 단어행렬은 변수의 개수가 관측치의 개수에 비해 크기 때문에 일반적인 주성분분석을 사용할 수 없다. 이와 같은 문제를 해결할 수 있는 하나의 방법으로 본 논문에서는 SVD의 사용을 제안한다. SVD에서 차원이 $(n\times p)$인 행렬 A는 다음과 같이 표현된다[21-22].

$$U^{'}AV=\begin{pmatrix}D\\0\end{pmatrix} \tag{1}$$

여기서 U는 차원이 $(n\times n)$인 직교행렬(orthogonal matrix)이고 V는 차원이 $(p\times p)$인 직교행렬이다. D는 $diag(\Delta_1,\Delta_2,...,\Delta_p)$인 대각행렬(diagonal matrix)이다. D에서 Δ_i는 행렬 A의 비정칙값(singular values)이고 이 값은 A′A의 고유값(eigen value)의 양의 제곱근이다. 이를 통하여 다음과 같이 각 특허문서를 주성분(principal component)에 의한 선형결합(linear combination)으로 나타낼 수 있다[23].

$$Y_i=l_i^{'}X=l_{1i}X_1+l_{2i}X_2+...+l_{ki}X_k \tag{2}$$

여기서 X_i는 원래 데이터의 i번째 변수이고, l_i는 이에 대응되는 인자 적재(loading)값이다. 위의 식을 이용하여 대부분의 셀이 0인 단어행렬로부터 모든 셀이 연속형(continuous)값으로 바뀐 SVD-주성분행렬을 얻게 된다. 본 논문에서 최적의 군집 수 결정을 위한측도로서 AIC, BIC, 그리고 실루엣(Silhouette)을 이용한 결합(voting)전략을 사용한다. 1980년대 후반에 Rousseeuw에 의해 제안된 실루엣은 주어진 데이터의 군집화 결과에 대한 성능평가를 위한 측도로써 각 개체가 자신이 속한 군집에 얼마나 잘 포함되고 있는지를 시각적 표현(graphical representation)으로 제공한다[24]. 각 데이터 i에 대하여 $a(i)$는 i와 같은 군집에 속한 다른 모든 데이터와 i의 평균거리(average dissimilarity)를 나타낸다. $a(i)$값이 작을수록 군집화는 잘 이루어진 것으로 판단한다. 다음으로 i와 다른 한 개의 군집에 속한 데이터와의 평균거리를 계산한다. i가 속하지 않은 모든 군집들과의 평균거리를 반복적으로 구한 후에 i와 각 군집간의 가장 작은 평균거리를 $b(i)$라 하면 다음과 같이 $s(i)$가 정의된다.

$$s(i)=\frac{b(i)-a(i)}{\max a(i),b(i)} \tag{3}$$

여기서 $s(i)$는 다음과 같은 범위의 값을 갖는다.

$$-1\le s(i)\le 1 \tag{4}$$

$s(i)$값이 1에 가까울수록 군집화는 잘 되었다고 평가할 수 있다. 이 값이 -1에 가깝게 되면 i는 이웃군집(neighbouring cluster)으로 이동해야 한다. 이웃군집이란 $b(i)$의 조건을 만족하는 군집이다. $s(i)$가 0에 가까우면 i는 2개의 군집경계에 있게 된다. 따라서 평균 $s(i)$값인 실루엣이 가장 큰 경우의 군집화가 가장 우수하며 이 때 사용된 군집수가 주어진 데이터집합의 최적 군집수가 된다. 다음으로 고려되는 평가측도는 AIC(Akaike′s information criterion)이다. 다음 식은 AIC에 대한 정의이다[20-21].

$$AIC=nln\left(\frac{ESS_p}{n}\right)+2(p+1) \tag{5}$$

위 식에서 p는 독립변수의 개수이고 n은 데이터의 크기, 그리고 ESS_p는 p개의 독립변수를 갖는 모형의 오차제곱합(error sum of squares)이다. 가장 작은 AIC 값에서 최적의 군집수가 결정된다. 본 논문에서는 AIC와 함께 다음의 BIC(Bayesian information criterion)도 최적군집화를 위한 측도로서 고려한다[20-21].

$$BIC=-2\ln(L)+k\ln(n) \tag{6}$$

여기서 n은 데이터의 크기이고 k는 추정된 자유모수(free parameters)의 개수이다. L은 추정모형을 위한 가능도함수(likelihood function)의 최대값이다. 본 논문에서는 실루엣, AIC, 그리고 BIC를 결합(voting)하여 최적의 군집수를 결정하고 이를 이용하여 특허데이터를 군집화하는 앙상블모형을 제안한다.

4. 실험 및 결과

제안방법의 성능을 평가하기 위한 실험을 위하여 다음의 키워드 검색식(keyword equation)을 이용하여 KIPRIS (Korea Intellectual Property Rights Information Service)에서 제공하는 미국의 특허데이터를 사용하였다.

$$AB=(Data*Mining)+(Knowledge*Discovery) \tag{7}$$

위 식에서 AB는 요약(abstract)을 나타내고 '*'와 '+'는 각각 'and'와 'or'연산을 나타낸다. 실험에 사용될 특허문서는 데이터마이닝의 지식추출에 관한 기술을 포함한다. 1987년 데이터마이닝 기술에 관한 최초 특허출원이 이루어진 이후 현재까지 등록된 전체 특허수는 모두 99개였다. 본 실험에서는 데이터마이닝에 관련된 세부기술들 중에서 아직 개발이 제대로 이루어지지 않고 있는 공백기술을 예측하려고 한다. 다음은 실험에 사용될 특허자료에 대한 요약이다.

표 1. 검색된 특허데이터
Table 1. Retrieved patent data

Data		출원연도	등록 특허건수
전체	학습	1987 - 2002	67
	검증	2003 - 2007	32

본 논문에서는 전체 99개의 특허문서를 학습데이터(training data)와 검증데이터(test data)로 나누었다. 일반

적인 기계학습에서 제공되는 기준에 따라 전체데이터의 2/3는 학습(training)을 위하여 그리고 나머지 1/3은 검증(test)을 위하여 나누었다[25]. 우선 USPTO(United States Patent and Trademark Office)[26]로부터 KIPRIS를 통해 얻은 99개의 특허데이터 중에서 67개의 학습데이터에 대한 텍스트 마이닝(text mining)[27]과 같은 전처리 과정을 통하여 다음과 같은 단어행렬을 구하였다. 본 실험의 분석을 위하여 통계계산을 위한 소프트웨어인 R을 이용하였다[28].

$$T_{67\times 1694} = \begin{pmatrix} t_{1,1}, & t_{1,2}, & \cdots, & t_{1,1694} \\ t_{2,1}, & t_{2,2}, & \cdots, & t_{2,1694} \\ \vdots & \vdots & \ddots & \vdots \\ t_{67,1}, & t_{67,2}, & \cdots, & t_{67,1694} \end{pmatrix} \quad (8)$$

단어행렬 T는 67개의 특허문서와 1694개의 단어로 이루어진 행렬구조를 가지고 있다. T에 속한 t_{ij}는 i번째 문서에 나타난 j번째 단어의 빈도수를 나타낸다. 관측치의 수(67)보다 변수의 수(1694)가 훨씬 크기 때문에 일반적인 주성분분석의 수행은 불가능함을 알 수 있다. 그러므로 본 논문에서는 SVD를 이용하여 다음과 같은 SVD-주성분행렬을 구하였다.

$$SP_{67\times 67} = \begin{pmatrix} sp_{1,1}, & sp_{1,2}, & \cdots, & sp_{1,67} \\ sp_{2,1}, & sp_{2,2}, & \cdots, & sp_{2,67} \\ \vdots & \vdots & \ddots & \vdots \\ sp_{67,1}, & sp_{67,2}, & \cdots, & sp_{67,67} \end{pmatrix} \quad (9)$$

SVD를 이용한 주성분분석으로부터 얻을 수 있는 최대 주성분의 수는 관측치의 개수 만큼이다[23]. 이는 앞의 식(2)에서의 lX에 해당한다. 위 행렬에서 sp_{ij}는 i번째 문서의 j번째 SVD-주성분점수를 나타낸다. 공백기술예측을 위하여 전체문서를 SVD-주성분점수에 의해 군집화를 수행한다. 군집화를 위하여 가장 먼저 결정해야 하는 것은 군집수이다. 본 논문에서는 AIC, BIC, 그리고 실루엣의 3가지 측도를 이용하여 결정한다. 먼저 실루엣에 의한 분석결과는 다음과 같다. 가장 큰 실루엣값에서 최적군집수를 결정한다.

표 2. 군집수 결정을 위한 실루엣 값
Table 2. Silhouette for determining number of clusters

군집수	사용된 주성분 수 (상위기준)				
	3	5	10	20	67
4	0.42	0.39	0.26	0.17	0.17
5	0.36	0.27	0.29	0.21	0.17
6	0.41	0.31	0.31	0.23	0.19
7	0.42	0.33	0.33	0.25	0.20
8	0.45	0.37	0.34	0.26	0.16
9	0.47	0.38	0.36	0.26	0.17
10	0.41	0.36	0.35	0.28	0.16

실루엣측도에 의한 최적군집수는 사용된 주성분의 수가 3개와 10개 일 때는 9, 5개, 20개 일 때는 각각 4와 10이었다. 사용할 수 있는 모든 주성분수인 67개일 때는 최적군집수를 7로 결정할 수 있다. 다음으로 BIC 측도에 의한 분석결과는 다음과 같다. 가장 작은 BIC 값에서 최적의 군집수가 결정된다.

표 3. 군집수 결정을 위한 BIC 값
Table 3. BIC for determining number of clusters

군집수	사용된 주성분 수 (상위기준)				
	3	5	10	20	67
4	151.3	282.9	613.6	1350.0	4634.0
5	161.7	301.5	663.4	1474.8	5069.6
6	175.8	323.4	720.5	1606.3	5522.0
7	196.5	352.5	780.0	1739.2	6007.3
8	218.6	385.8	841.9	1872.6	6494.2
9	240.9	420.9	906.5	2006.8	6984.3
10	263.9	456.4	973.7	2143.4	7479.1

위 결과를 통하여 BIC에 의한 최적군집수는 모든 주성분수에서 4로 결정되었다. 최적 군집수 결정을 위한 마지막 측도인 AIC에 의한 분석결과는 다음 표와 같다. 이 측도도 BIC와 마찬가지로 가장 작은 값일 때 최적군집수가 결정된다.

표 4. 군집수 결정을 위한 AIC 값
Table 4. AIC for determining # of clusters

군집수	사용된 주성분 수 (상위기준)				
	3	5	10	20	67
4	98.4	194.7	437.2	997.3	3452.3
5	95.5	191.3	442.9	1033.8	3592.4
6	96.4	191.1	456.0	1077.2	3749.4
7	103.9	198.2	471.3	1121.9	3939.3
8	112.8	209.5	489.1	1167.1	4130.8
9	121.8	222.5	509.7	1213.1	4325.4
10	131.7	236.0	532.8	1261.5	4524.8

주성분수가 3개와 5개 일 때 AIC에 의한 최적군집수는 각각 5와 6이다. 나머지 주성분개수에는 모두 최적군집수가 4로 결정된다.

실루엣, BIC, 그리고 AIC의 결과를 결합하여 최적의 군집수를 결정하기 위한 요약된 표는 다음과 같다.

표 5. 각 군집화 측도에서 나타난 최적군집수
Table 5. Optimal number of clusters

군집화 측도	사용된 주성분 수 (상위기준)				
	3	5	10	20	67
Silhouette	8	4	9	10	7
BIC	4	4	4	4	4
AIC	5	6	4	4	4

따라서 최적군집수 결정을 위한 후보군집수의 선택비율(voting rate)은 다음과 같다.

표 6. 최적 군집수를 위한 후보군집수의 선택비율
Table 6. Voting rate of candidate for optimal clustering

후보군집수	voting rate (%)	
4	9/15	60.0%
5	1/15	6.7%
6	1/15	6.7%
7	1/15	6.7%
8	1/15	6.7%
9	1/15	6.7%
10	1/15	6.7%
합계	15/15	100%

위 결과에서 가장 큰 선택값(voting rate)을 갖는 후보군집수는 4이다. 그러므로 본 실험에 사용된 특허데이터의 최적군집수는 4로 결정할 수 있다. 이 결과를 이용하여 k값이 4인 k-means 군집화[29-30]를 수행한다. 다음은 K-means 군집화 결과이다.

표 7. K-평균 군집화 결과
Table 7. K-means clustering result

군집	소속특허 (일련번호)	특허수	비율(%)
1	1, 3, 10, ..., 65, 66, 67	39	58.21
2	34, 36, 37, ..., 58, 59, 63	9	13.43
3	31	1	1.49
4	2, 4, 5, ..., 32, 49, 54	18	26.87

위의 분석결과를 통하여 군집1과 군집4는 이미 데이터마이닝 관련 세부기술에 대한 연구, 개발이 활발하게 진행되고 있는 분야임을 알 수 있다. 이에 비해 군집3은 단지 1개의 특허기술을 포함하고 있다. 군집3에 속한 구체적인 등록특허의 제목은 다음과 같다.

'Self-contained mapping and positioning system utilizing point cloud data'

이 기술은 다른 모든 데이터마이닝 관련 기술과는 동떨어진 내용을 포함하고 있다. 클라우드 데이터(cloud data)의 효율적인 사용을 위한 데이터 마이닝 관리시스템에 관한 것이다. 즉, 데이터 마이닝 기술에 대한 세부기술이라기 보다는 클라우딩 컴퓨팅기반 서비스를 위한 세부기술에 해당되기 때문에 데이터 마이닝 기술분야와는 거리가 있는 이상치(oulier) 기술로 판정할 수 있다. 따라서 본 논문에서는 군집2에 속하는 기술을 공백기술로 판정하였다.

군집2가 나타내는 세부기술을 정의하기 위하여 이 군집에 속한 특허들로부터 대표적인 키워드를 추출하였다. 상위 5개의 키워드는 'schema', 'predefined', 'object', 'block', 그리고 'control'이었다. 이를 통하여 군집2가 대표하는 세부기술은 데이터마이닝 응용시스템을 구축하기 위한 마이닝 작업에 대한 디자인기술로 정의하였다. 즉, 최종사용자가 데이터마이닝 결과를 이용하여 최적의 의사결정을 할 수 있도록 초기에 마이닝시스템의 구축에 대한 디자인 기술이 필요할 것이고 이에 대한 기술연구는 기존의 분석기법 위주의 데이터 마이닝 연구에 비해 아직 제대로 이루어지고 있지 않음을 알 수 있다. 그러므로 이에 대한 연구, 개발이 필요하다고 판단하였다. 제안된 기법의 성능을 평가하기 위하여 2003년 이후의 특허에 대하여 군집2에 속한 특허들의 빈도를 계산하였다. 총 32개의 특허들 중에서 군집2에 속한 특허의 개수는 8개였다. 이는 테스트 데이터의 25%를 차지하는 것으로서 학습 데이터에서의 13%에 비해 약 2배의 증가를 확인 할 수 있었다. 이를 통하여 군집2의 기술이 공백기술이라는 본 논문의 판정결과의 성능을 확인하여 주었다.

5. 결 론

본 논문에서는 특허문서를 이용한 공백기술예측에 대한 방법을 제안하여 이를 지능형시스템 분야에 적용하였다. 학교. 기업, 그리고 정부의 연구기관들이 서로 중복된 연구를 수행할 수도 있고, 불필요한 기술에 대하여 시간과 비용을 낭비할 수도 있다. 본 연구에서 제안된 방법을 통하여 특허데이터를 분석하고, 공백기술을 예측하고 이를 바탕으로 향후 지능형시스템 분야에서 필요하게 될 세부기술에 대한 계획적인 연구, 개발이 이루어 질 수 있을 것이다. 물론 본 연구에서 제시한 방법은 지능형시스템분야 이외의 다양한 기술분야에도 적용될 수 있다. 하지만 본 연구의 실험결과와 같이 2개의 세부 공백기술 군집을 예측했지만 이 중에 하나는 앞으로 증가할 공백기술이 아닌 것으로 확인되었다. 이와 같은 문제를 해결하기 위해서 좀 더 정교한 분석기법에 대한 연구와 함께 분석결과에 대한 해석에 대한 과정에서 기술분야에 전문가 집단이 폭넓게 참여해야 할 것이다.

본 논문에서는 지능형시스템에 대한 연구결과로서 지금까지 출원, 등록된 미국특허만을 이용하였다. USPTO의 특허데이터베이스에는 미국의 학교, 기업 및 정부의 모든 연구기관의 연구결과와 미국시장을 겨냥한 전 세계의 연구기관으로부터 출원된 결과를 포함하고 있지만 지능형시스템에 대한 모든 연구결과를 포함한 것은 아니다. 앞으로 미국특허청 뿐만 아니라 유럽특허청, 일본특허청, 그리고 대한민국특허청으로부터의 특허데이터와 IEEE를 포함한 주요 논문지에 발표된 논문까지 포함한 광범위한 연구개발 결과를 이용한 기술예측이 이루어진다면 지능형시스템 기술에 대한 좀 더 정확한 기술예측 결과가 제공될 수 있을 것이다. 이 부분과 더 발전된 새로운 기술예측 방법에 대한 연구는 향후 과제로 남긴다.

참 고 문 헌

[1] *법률지식정보시스템*, likms.assembly.go.kr
[2] 나까야마 노부히로, *특허법*, 법문사, 2001.
[3] 제대식, 이은철, 윤국섭 역, *지식경영과 특허전략*, 세종서적, 2000.
[4] 황종환, *특허법*, 한빛지적소유권센터, 2001.
[5] 남영준, 정의섭 "인용정보를 이용한 신 특허지수 개발에 관한 연구," *정보관리학회지*, vol. 23, no. 1, pp. 221-241, 2006.
[6] 유선희, "특허인용 분석을 통한 기술수명예측모델 개발에 관한 연구," *정보관리연구*, vol. 35, no. 1, pp. 93-112, 2004.
[7] 전성해, 박상성, 신영근, 장동식, 정호석, "자기조직화지도와 매트릭스분석을 이용한 특허분석시스템의 공백기술 예측," *한국콘텐츠학회논문지*, vol. 10, no. 2, pp. 462-480, 2010.

[8] 전성해, 특허분석을 이용한 지능형시스템의 기술예측, *한국지능시스템학회 논문지*, vol. 21, no. 1, pp. 100-105, 2011.
[9] B. Yoon, Y. Park, "Development of New Technology Fore casting Algorithm: Hybrid Approach for Morphology Analysis and Conjoint Analysis of Patent Information," *IEEE Transactions on Engineering Management*, vol. 54, no. 3, pp. 588-599, 2007.
[10] 특허청 정보기획팀, 한국발명진흥회 정보활용지원팀, *특허와 정보분석* (개정판), 성민, 2007.
[11] B. Yoon, S. Lee, "Patent analysis for technology forecasting: Sector-specific applications," *Proceeding of IEEE International Conference on Engineering Management*, pp. 1-5, 2008.
[12] *특허정보검색서비스*, www.kipris.or.kr
[13] 전성해, 엄대호, "특허와 통계학, 그 연결은?" *한국통계학회논문집*, vol. 17, no. 2, pp. 205-222, 2010.
[14] S. Lee, B. Yoon, Y. Park, "An approach to discovering new technology opportunities: Keyword-based patent map approach," *Technovation*, vol. 29, pp. 481-497, 2009.
[15] P. Wang, I. M. Cockburn, M. L. Puterman, " Analysis of Patent Data-A Mixed Poisson Regression Model Approach," *Journal of Business & Economic Statistics*, vol. 16, no. 1, pp. 27-41, 1998.
[16] J. Han, M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001.
[17] M. Fattori, G. Pedrazzi, R. Turra, "Text mining applied to patent mapping: a practical business case," *World Patent Information*, vol. 25, pp. 335 - 342, 2003.
[18] G. H. Golub, C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, vol. 14, no. 5, pp. 403-420, 1970.
[19] 오일석, *패턴인식*, 교보문고, 2008.
[20] J. F. Hair, B. Black, B. Babin, R. E. Anderson, *Multivariate Data Analysis*, Prentice Hall, 1992.
[21] R. A. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis*, Pearson, 2007.
[22] 김병천 역, *통계학을 위한 행렬대수학*, 자유아카데미, 2001.
[23] 강근석, 김충락, *회귀분석*, 교우사, 2005.
[24] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.
[25] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
[26] *USPTO* (United States Patent and Trademark Office), www.uspto.gov
[27] I. Feinerer, K. Hornik, D. Meyer, "Text Mining Infrastructure in R," *Journal of Statistical Software*, vol. 25, iss. 5, pp. 1-54, 2008.
[28] R Development Core Team, R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org, 2010.
[29] X. Chen, W. Yin, P. Tu, H. Zhang, "Weighted k-Means Algorithm Based Text Clustering," *Proceedings of International Symposium on Information Engineering and Electronic Commerce*, pp. 51-55, 2009.
[30] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," IEEE transaction on pattern analysis and machine intelligence, vol. 24, pp. 881 - 892, 2002.

저 자 소 개

전성해(Sunghae Jun)
1993년 : 인하대 통계학과(학사)
1996년 : 인하대 통계학과(이학석사)
2001년 : 인하대 통계학과(이학박사)
2007년 : 서강대학교 컴퓨터공학과 (공학박사)
2003년～현재 : 청주대학교 바이오정보통계학과 부교수

관심분야 : 기술경영, 인공지능, 데이터마이닝
Phone : 043-229-8205
Fax : 043-229-8432
E-mail : shjun@cju.ac.kr