KIPS Tr. Software and Data Eng.
Vol.3, No.9 pp.355~360 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.9.355

Time Series Analysis of Patent Keywords for Forecasting Emerging Technology

Jong-Chan Kim† · Joon-Hyuck Lee†† · Gab-Jo Kim††† · Sang-Sung Park†††† · Dong-Sick Jang†††††

ABSTRACT

Forecasting of emerging technology plays important roles in business strategy and R&D investment. There are various ways for technology forecasting including patent analysis. Qualitative analysis methods through experts' evaluations and opinions have been mainly used for technology forecasting using patents. However qualitative methods do not assure objectivity of analysis results and requires high cost and long time. To make up for the weaknesses, we are able to analyze patent data quantitatively and statistically by using text mining technique. In this paper, we suggest a new method of technology forecasting using text mining and ARIMA analysis.

Keywords : Technology Forecasting, Text Mining, Patent Data, ARIMA Analysis

특허 키워드 시계열 분석을 통한 부상 기술 예측

김 종 찬† · 이 준 혁†† · 김 갑 조††† · 박 상 성†††† · 장 동 식†††††

요 약

오늘날 국가와 기업의 연구 개발 투자 및 경영 정책 전략 수립에서 미래 부상 기술 예측은 매우 중요한 역할을 한다. 기술 예측을 위한 다양한 방법들이 사용되고 있으며 특허를 이용한 기술 예측 또한 활발히 진행되고 있다. 특허를 이용한 기술 예측에는 전문가들의 평가와 견해를 통한 정성적인 방법이 주로 사용되어 왔다. 정성적인 방법은 분석 결과의 객관성을 보장하지 못하고 분석에 많은 비용 및 시간이 요구된다. 이런 문제점을 보완하기 위해 최근에는 텍스트 마이닝을 이용한 특허 데이터의 정량적인 분석이 이루어지고 있다. 텍스트 마이닝 기법을 적용함으로써 특허 문서의 통계적 분석이 가능하다. 본 논문에서는 텍스트 마이닝과 ARIMA 분석을 이용한 기술 예측 방법을 제안한다.

키워드 : 기술 예측, 텍스트 마이닝, 특허 데이터, ARIMA 분석

1. 서 론

오늘날 국가와 기업들은 효과적인 연구 개발 투자와 경영 정책 수립을 위해 미래의 기술 동향 및 부상 기술 예측에 많은 노력을 기울이고 있다. 기술 예측을 위한 다양한 방법들이 사용되고 있으며 새로운 기술 예측 방법을 위한 많은 연구가 진행되고 있다. 그 중 특허 데이터를 이용한 기술 예측 방법도 활발히 진행되고 있다. 특허 데이터는 기술의 정보를 서지적 사항(출원번호, 출원인, 인용특허, IPC 코드 등)과 기술적 사항(발명의 명칭, 요약, 발명의 상세한 설명 등)으로 나누어 명확히 기록하고 있다[1]. 과거에는 특허의 기술적 정보를 분석하는데, 대부분 전문가의 정성적인 방법이 사용되어 왔다. 그러나 최근 텍스트 마이닝 기법을 통해 특허의 기술적 정보를 정량적인 방법으로 분석하여 기술을 예측하는 연구가 활발히 진행되고 있다[2]. 본 연구는 특허의 기술적 사항에서 텍스트 마이닝 기법을 통해 핵심 키워드를 추출하고 추출된 핵심 키워드의 연도별 출현빈도를 이용해 시계열 분석을 하였다. 이 분석 결과를 통해 미국 탄소 복합소재 분야의 부상 기술을 예측하였다.

※ 본 논문은 BK21 플러스 사업(고려대학교, 제조·물류분야에서의 빅데이터 운용 사업팀)으로 지원된 연구임.
※ 본 논문은 2012년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(한국연구재단-NRF-2010-0024163).
† 준 회 원 : 고려대학교 산업경영공학과 석·박사통합과정
†† 준 회 원 : 고려대학교 산업경영공학과 박사과정
††† 정 회 원 : 고려대학교 산업경영공학과 석·박사통합과정
†††† 정 회 원 : 고려대학교 산업경영공학부 연구교수
††††† 정 회 원 : 고려대학교 산업경영공학부 교수
Manuscript Received : July 4, 2014
First Revision : August 1, 2014
Accepted : August 1, 2014
* Corresponding Author : Dong-Sick Jang(jang@korea.ac.kr)

2. 이론적 배경

2.1 텍스트 마이닝

특허 데이터와 같은 문서들은 일반적으로 시계열 분석, 군집 분석과 같은 통계 분석 시 이용하는 수치형 데이터가 아닌 텍스트 형태의 비정형 데이터이기 때문에 기존의 통계 분석 방법을 적용하기에 어려움이 있다. 텍스트 형태의 데이터에 기존 데이터 마이닝 및 통계 기법을 이용하기 위해서는 비구조화 데이터들을 텍스트 마이닝을 통해 구조화 데이터로 변환해야 한다[3]. S. H. Jun(2009)는 특허 정보 분석을 위해 전처리와 변환 작업을 포함하는 효율적인 텍스트 마이닝 방법을 제안하였다[4].

2.2 ARIMA 분석

1) Auto-Regression(AR)

Auto-regressive model이란 시계열 자료(Y_t)가 그 시계열 자료의 과거 값들로써 설명된다는 모형이다. 다시 말하면, 회귀 모형의 독립 변수들이 그 시계열 자료의 과거 값들이라는 것이다. 즉, 다음 Auto-regressive model 식 1이라고 할 때,

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t,$$
$$\varepsilon_t \sim i.i.d.\, N(0, \sigma^2) \quad (1)$$

현재(t) 시점의 시계열 값(Y_t)은 바로 전기(t-1) 값으로부터 α_1만큼, 2시점 전인 (t-2)의 값으로부터 α_2만큼 영향을 받는다고 할 수 있다[5, 6].

2) 이동평균 방법(MA)

이동평균 방법이란 어느 시계열 자료에 대해 예측 시점(T)을 기준으로, 차기(T+1)의 예측 값을 시점 T에서 가지고 있는 과거 자료들의 평균 값으로 하는 방법이다. 예측 시점에서 평균을 구할 때 이전 시점의 자료들을 몇 개 사용하느냐에 따라 평균 값이 다르게 되므로 이동평균이라는 용어를 사용한다. 이동평균 방법에는 단순 이동평균, 중심화 이동평균, 선형 이동평균이 있다[5, 6].

3) 차분

Box-jenkins 모형을 비롯한 대부분의 시계열 분석 모형은 분석 대상이 정상적인 시계열 데이터라고 가정된 상태에서 만들어졌기 때문에 비정상적인 시계열 데이터는 차분(difference)을 통해 정상적인 시계열 데이터로 변환하는 것이 필요하다. 시계열 자료, $\{Y_t\}$가 있을 때 1차 차분은

$$\triangle Y_t = Y_t - Y_{t-1} = Y_t - BY_t = (1-B)Y_t \quad (2)$$

으로 표현되며 1차 차분의 결과인 $\triangle Y_t$가 정상적 시계열이 아니라면 $\triangle Y_t$를 다시 한 번 차분하는 2차 차분을 수행할 수 있다. 즉, 2차 차분은 아래의 식 3과 같다.

$$\begin{aligned} \triangle^2 Y_t &= \triangle Y_t - \triangle Y_{t-1} \quad (3) \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2} \\ &= (1 - 2B + B^2) Y_t \\ &= (1-B)^2 Y_t \end{aligned}$$

대개의 많은 비정상적 시계열 자료들은 1차 차분 또는 2차 차분으로 정상적 시계열이 된다.

4) ARIMA 모형

비정상적인 시계열 데이터의 차분(d)를 고려하여 AR(p) 모형과 MA(p) 모형이 합쳐진 모형이 ARIMA(p, d, q) 모형이다. ARIMA 모형은 아래와 같이 표현된다[5, 6].

$$\begin{aligned} Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t \quad (4) \\ - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - \cdots - \beta_p \varepsilon_{t-p} \end{aligned}$$

3. 선행 연구

기존의 텍스트 마이닝을 이용한 특허 정보 분석으로는 문서-단어 행렬을 추출하고 문서를 군집하여 기술을 예측하는 방법이 있다. Tseng(2009) 외 2명은 텍스트 마이닝과 topic clustering을 이용하여 전계 방출 디스플레이 기술 분야의 트렌드를 알아보았다. 또 Jun(2011)는 k-means 알고리즘을 이용하여 지능형 시스템의 공백 기술을 예측하였고 Kim(2012)은 CLARA 알고리즘을 이용하여 OLED 기술 분야의 공백 기술을 예측하였다[7, 8, 9]. 문서 군집화를 이용한 기술 예측 방법은 비슷한 특징을 갖는 특허 문서들의 군집을 통해 기술을 정의하기 때문에 개개의 키워드들이 갖고 있는 의미를 고려하여 기술을 정의하는 데 어려움이 있다. 위와 같은 문서 군집을 통한 분석뿐만 아니라 키워드 기반의 분석을 수행한 연구도 있다. Yoon과 Park은 키워드 기반의 형태 분석을 이용해 TFT-LCD 기술 분야의 기술 기회 분석을 수행하였고 Lee 외 2명은 키워드 기반의 Patent Map을 이용해 PDA technology 기술 분야의 공백 기술을 찾아냈다[10, 11]. 텍스트 마이닝을 이용해 특허 문서를 분석하여 미래 기술을 예측한 기존의 연구들은 예측에 시간적인 개념에 중점을 두고 고려하지 않고 있었다. 예측을 위한 시계열 분석으로는 회귀 분석,

지수평활, 이동평균, ARIMA 분석 등 여러 가지 방법들이 있다. Ediger와 Akar는 ARIMA 분석을 이용해 터키의 1차 에너지 수요 예측을 했다[12]. 본 연구는 특허 문서에서 핵심 키워드를 추출하여 키워드 기반의 분석을 수행하고 핵심 키워드의 연도별 출현빈도를 시계열 자료로 하여 시계열 분석 방법 중 하나인 ARIMA을 이용한 분석을 하였다.

4. 제안된 연구 방법

먼저 WIP SON, KIPRIS와 같은 특허 데이터베이스를 통해 예측을 하고자 하는 기술 분야의 특허 데이터를 수집한다[13, 14]. 수집된 특허 데이터에서 각 기술에 대한 설명을 포함하는 발명의 명칭, 요약, 대표 청구항을 추출한다. 추출된 데이터는 텍스트 마이닝 기법을 이용해 불용어, 공백 등을 제거하는 전처리 과정을 거쳐 문서-단어 행렬을 생성한다. 생성된 문서-단어 행렬에서 단어의 출현빈도가 200개 미만인 단어를 제거한 후 문서-단어 행렬에 출현빈도 대신 TF-IDF 가중치를 적용 하여 핵심 키워드를 추출한다. 추출 된 핵심 키워드의 연도별 출현빈도를 조사하고 이를 시계열 자료로 하여 ARIMA 분석을 이용한 예측을 수행한다. 그리고 ARIMA 분석을 통해 2010년도의 각 키워드 출현빈도의 예측치를 계산한다. 계산된 2010년도의 예측치와 2000년부터 2009년도까지의 평균치를 비교하여 각 핵심 키워드의 부상도를 계산한다. 계산된 부상도가 0 이상인 핵심 키워드를 선정하여 부상 기술을 예측할 수 있다. 본 논문에서 제안하는 연구 프로세스는 아래 그림 1과 같다.

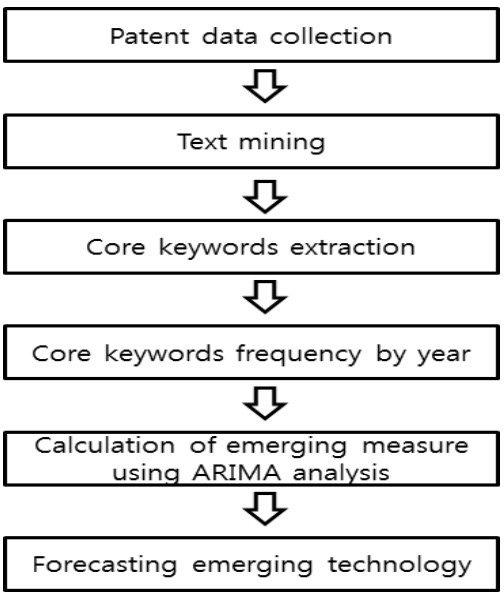

Fig. 1. Research Methodology

5. 실험 및 결과

우리는 제안된 방법을 이용한 실험을 통해 탄소 복합소재 분야의 부상 기술을 도출하였다. 먼저 특허 데이터베이스 중 하나인 WIPS ON을 통해 미국에서 출원된 탄소 복합소재 기술 특허를 2000년부터 2009년까지 출원 연도별로 수집하였다. 수집된 특허 문서의 수는 아래 표 1과 같다.

Table 1. The number of patent applications by year

Year	Number of patents	Year	Number of patents
2000	25	2005	63
2001	44	2006	86
2002	71	2007	71
2003	93	2008	69
2004	82	2009	82

키워드를 이용한 정량 분석을 위해 수집된 특허 데이터 집합의 다양한 정보 중에서 중요한 기술적 설명을 포함하고 있는 발명의 명칭, 요약, 대표 청구항을 추출하였다. 우리는 추출된 데이터를 텍스트 마이닝 기법을 이용한 전처리 과정을 통해 불용어, 공백 등을 제거하고 분석이 가능한 문서-단어 행렬로 변환하였다. 탄소 복합소재 분야의 기술을 대표하는 핵심 키워드를 도출하기 위해 문서-단어 행렬 안에 출현빈도 대신 TF-IDF 가중치를 부여하고 핵심 키워드와 노이즈를 판별하기 위한 기준으로 TF-IDF 임계치를 이용했다. 최적의 TF-IDF 임계치를 구하기 위해 실험 데이터 중 임의로 추출한 30개의 표본을 통해 각 표본이 어느 수치의 임계치에서 핵심 키워드와 노이즈를 잘 구분하는지 정성적인 방법으로 성능을 검증하였다. 그 결과 0.15가 최적의 TF-IDF 임계치로 도출되었고 이를 이용해 0.15 이상의 TF-IDF 가중치를 갖고 출현빈도가 200개 이상인 핵심 키워드 21개를 아래 표 2와 같이 선정하였다.

분석 결과로 도출된 표 2의 핵심 키워드들은 수집된 탄소 복합소재 특허 문서들의 기술 특징을 대표하는 단어들이라고 할 수 있다. 우리는 핵심 키워드들의 출현빈도가 증가할수록 해당 키워드와 관련된 기술들이 발전할 것이라 예상할 수 있다. 그러므로 우리는 2000년부터 2009년까지 연도별 출현빈도를 이용한 ARIMA 분석을 통해 부상 기술을 예측하였다. 이를 통해 2010년의 핵심 키워드의 출현빈도를 예측하고 과거 데이터의 평균과 비교하여 부상도가 0 이상인, 즉 출현빈도가 증가할 것이라 예상되는 부상 기술 키워드로 outline, ground, rolled, microporous, laminate, machine, permeable, durable, pumps, periodic, outer, pinstock을 선택하였다.

Table 2. Core keywords selected

Core keywords	Appearance frequency	Maximum TF-IDF
periodic	851	0.1982566
outer	459	0.2722689
responsive	438	0.2705839
pinstock	374	0.3070386
outline	275	0.2793916
impermeable	272	0.3520837
mat	265	0.1994743
rolled	257	0.3440263
ground	257	0.2839716
union	249	0.4084122
microporous	241	0.4085805
sail-attaching	240	0.3604917
microstructures	238	0.3391898
solution	238	0.2172179
volumes	237	0.1715892
laminate	233	0.1734323
machine	231	0.2606183
permeable	230	0.261094
pumps	212	0.179552
leg	204	0.3591737
durable	200	0.4048422

Table 3. Emerging value by ARIMA analysis

Core keywords	Existing value	Predictive value	Emerging value	Validation data
impermeable	27.2	22.45909	-4.74091	0
responsive	43.8	41.58088	-2.21912	1
mat	26.5	24.8207	-1.6793	11
sail-attaching	24	22.96644	-1.03356	0
leg	20.4	19.8825	-0.5175	0
volumes	23.7	23.31235	-0.38765	9
solution	23.8	23.42685	-0.37315	28
union	24.9	24.9	0	0
laminate	23.3	23.49494	0.19494	24
rolled	25.37	26.20629	0.50629	10
ground	25.7	26.54862	0.84862	4
durable	20	21.64323	1.64323	1
microstructures	23.8	25.5762	1.7762	0
pinstock	37.4	39.30162	1.90162	0
permeable	23	26.19135	3.19135	6
microporous	24.1	27.42865	3.32865	1
pumps	21.2	24.56035	3.36035	1
machine	23.1	28.2565	5.1565	12
outer	45.9	51.7005	5.8005	8
outline	27.5	39.30024	11.80024	0
periodic	85.1	110.01216	24.91216	6

따라서 이 부상 기술 키워드와 관련이 있는 기술들의 연구개발이 필요할 것으로 보인다. 예측 결과의 성능을 검증하기 위해 2010년의 실제 핵심 키워드 출현빈도와 비교했을 때 solution과 laminate를 제외한 모든 핵심 키워드의 출현빈도가 감소하였다. 즉, 실제 부상 기술 키워드는 solution과 laminate였다. 그러므로 우리는 실험을 통해 2개의 부상 기술 키워드 중 하나인 laminate가 부상 기술 키워드인 것을 예측할 수 있었다. ARIMA 분석이 아닌 지수평활법을 이용한 예측에서는 부상 기술 키워드로 periodic, pinstock, microstructures, solution, leg, machine, outer가 도출되었다[15]. 지수평활법을 이용한 분석은 laminate가 부상 기술 키워드라는 것을 예측하지 못했지만 solution이 부상 기술 키워드라는 것을 예측했다. 이 결과를 통해 우리는 다양한 시계열 분석 방법을 적용하여 예측 정확도를 높일 것이라 예상할 수 있다. 다음 표 3은 계산된 부상도와 실제 2010년도의 실제 출현빈도를 나타낸다.

6. 결 론

본 논문은 탄소섬유 복합소재 분야의 기술 예측을 위한 특허 정보 분석 방법으로 텍스트 마이닝 기법과 ARIMA 분석을 이용한 방법을 제안하였다. 수집된 특허 문서의 2,748개의 단어 중에서 TF-IDF 가중치와 출현빈도를 이용하여 핵심 키워드를 선정하였고 ARIMA 분석을 통해 얻은 예측치를 이용해 각 핵심 키워드의 부상도를 계산하였다. 이를 통해 부상도가 높은 12개의 단어(outline, ground, rolled, microporous, laminate, machine, permeable, durable, pumps, periodic, outer, pinstock)를 부상 기술 키워드로 선정하였다. 이 부상 기술 키워드와 관련된 기술이 탄소섬유 복합소재 분야의 부상 기술일 것으로 예측하였고 실제데이터를 통해 검증한 결과 laminate와 관련된 기술이 부상 기술이라는 것을 확인했다. 향후 연구에서는 예측의 정확도를 높이기 위해 ARIMA 분석 외에 다른 시계열 분석 방법의 적용이 필요하

고 텍스트 형태의 데이터를 시계열모형에 적용하기 위한 효율적인 전처리 방법에 대한 연구가 필요하다.

References

[1] Korean Intellectual Property Office, Korean Invention Promotion Association, “Patent and information analysis (for researchers),” Kyungsung Books, pp.302-372, 2009.

[2] B. U. 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.

[3] R. Feldman and J. Sanger, The text mining hand book: advanced approaches in analyzing unstructured data, Cambridge university press, pp.1-13, 2007.

[4] S. H. Jun, “An Efficient Text mining for Patent Information Analysis,” *Proceedings of KIIS Spring Conference*, Vol.19, No.1, pp.255-257, 2009.

[5] J. D. Hamilton, Time series analysis, Princeton university press, pp.25-142, 1994.

[6] E. A. Elsayed and T. O. Boucher, Analysis and control of production systems, Prentice Hall, pp.7-61, 1993.

[7] Y. H. Tseng, C.J. Lin, and Y. I. Lin, “Text mining technique for patent analysis,” *Information processing and management*, Vol.43, No.5, pp.1216-1257, 2009.

[8] S. H. Jun, “Technology forecasting of intelligent systems using patent analysis,” *Journal of Korean institute of intelligent systems*, Vol.21, No.1, pp.100-105, 2011.

[9] Y. S. Kim, S. S Park, and D. S. Jang, “Patent data analysis using CLARA algorithm: OLED technology,” *Journal of Korea institute of information technology*, Vol.10, No.6, pp.161-170, 2012.

[10] B. U. Yoon and Y. T. Park, “A systematic approach for identifying technology opportunities: keywords-based morphology analysis,” *Technologycal forecasting and social change*, Vol.72, No.2, pp.145-160, 2005.

[11] S. J. Lee, B. U. Yoon, and Y. T. Park, “An approach to discovering new technology opportunities: keywords-based patent map approach,” *Technovation,* Vol.29, No.6-7, pp. 481-497, 2009.

[12] V. S. Ediger and S. Akar, “ARIMA forecasting of primary energy demand by fuel in Turkey”, *Energy Policy,* Vol.35, No.3, pp.1701-1708, 2007.

[13] Wips on [internet], http://www.wipson.com/

[14] KIPRIS [internet], http://www.kpris.or.kr/

[15] J. C. Kim, J. H. Lee, G. J. Kim, S. S. Park, and D. S. Jang, “Time series analysis of patent keywords for forecasting emerging technology,” *The 2014 spring conference of the KIPS*, Vol.21, No.1, pp.650-652, 2014.

김 종 찬

e-mail : ourjongchan@korea.ac.kr

2013년 청주대학교 통계학과(학사)

2013년～현 재 고려대학교 산업경영공학과 석・박사통합과정

관심분야 : Technology Forecasting & Statistics

이 준 혁

e-mail : iguana751@korea.ac.kr

2012년 한국항공대학교 정보통신공학과 (학사)

2013년 고려대학교 산업경영공학과(석사)

2014년～현 재 고려대학교 산업경영공학과 박사과정

관심분야 : Predicting Firm Performance & Management of Technology

김 갑 조

e-mail : kkjjo@korea.ac.kr

2011년 숭실대학교 산업정보시스템공학과 (학사)

2011년～현 재 고려대학교 산업경영공학과 석・박사통합과정

관심분야 : Technology Forecasting & Data Mining

박 상 성

e-mail : hanyul@korea.ac.kr

2006년 고려대학교 산업경영공학과(박사)

2007년～현 재 고려대학교 산업경영공학부 연구교수

관심분야 : Management of Technology & Patent Analysis

장 동 식

e-mail : jang@korea.ac.kr

1988년 Texas A&M University 산업시스템 공학과(박사)

1989년~현 재 고려대학교 산업경영공학부 교수

관심분야 : Management of Patent & Strategic of Management