특허 키워드 시계열분석을 통한 부상기술 예측

김종찬, 이준혁, 김갑조, 박상성, 장동식
고려대학교 산업경영공학과
e-mail : ourjongchan@korea.ac.kr

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
Dept. of Industrial Management Engineering, Korea University

요 약

국가와 기업의 연구개발투자 및 경영정책 전략 수립에서 미래 부상기술 예측은 매우 중요한 역할을 한다. 기술예측을 위한 다양한 방법들이 사용되고 있으며 특허를 이용한 기술예측 또한 활발히 진행되고 있다. 최근에는 텍스트마이닝을 이용해 특허데이터의 정량적인 분석이 이루어지고 있다. 본 논문에서는 텍스트마이닝과 지수평활법을 이용한 기술예측 방법을 제안한다.

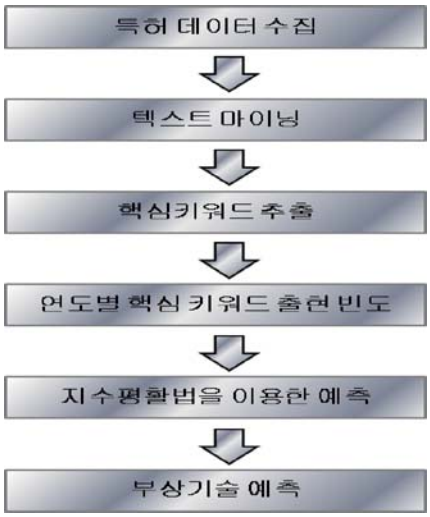
1. 서론

오늘날 국가와 기업들은 연구개발투자전략과 경영정책 수립을 위해 미래의 기술동향 및 부상기술 예측에 많은 노력을 기울이고 있다. 이러한 기술예측을 위해 다양한 방법들이 사용되고 있으며 새로운 기술예측 방법을 위한 많은 연구가 진행되고 있다. 그 중 특허데이터를 이용한 기술예측방법도 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 텍스트마이닝 기법을 통해 특허의 기술적 정보를 정량적인 방법으로 분석하여 기술을 예측하는 연구가 활발히 진행되고 있다[2]. 본 연구는 특허의 기술적 사항에서 텍스트마이닝 기법을 통해 핵심키워드를 추출하고 추출된 핵심키워드의 연도별 출현빈도를 이용해 시계열분석을 하였다. 이 분석결과를 통해 미국 탄소복합소재분야의 부상기술을 예측하였다.

2. 선행연구

기존의 텍스트마이닝을 이용한 특허정보분석으로는 문서-단어 행렬을 추출하고 문서를 군집하여 기술을 예측하는 방법이 있다. 전성해(2011)는 K-means 알고리즘을 이용하여 지능형시스템의 공백기술을 예측하였고 김요섭(2012)은 CLARA 알고리즘을 이용하여 OLED 기술분야의 공백기술을 예측하였다[3,4]. 문서군집화를 이용한 기술예측방법뿐 아니라 키워드 기반의 분석방법도 있다. 최진호 외 2 명은 텍스트마이닝과 키워드 네트워크분석을 이용해 LED 분야의 기술을 예측하였다[5]. 본 연구는 특허문서에서 핵심키워드를 추출하고 핵심키워드의 연도별 빈도수를 시계열 자료로 하여 지수평활(Exponential smoothing)법[6]을 이용한 분석을하였다. 이를 통해 미국의 탄소복합소재 기술을 예측하고자 한다.

3. 제안된 연구 방법



(그림 1) 연구 프로세스

WIPSON, KIPRIS 와 같은 특허 데이터베이스를 통해 예측을 하고자 하는 기술분야의 특허데이터를 수집한다[7, 8]. 수집 된 특허데이터를 텍스트마이닝 기법을 이용해 불용어, 공백 등을 제거하는 전처리 과정을 거쳐 문서-단어 행렬을 생성한다. 단어의 빈도수가 200 개 미만인 단어를 제거하고 TF-IDF 가중치를이용하여 핵심 키워드를 추출한다. 추출된 핵심 키워드의 연도별 출현빈도를 시계열 데이터로 하여 지수평활법을 이용한 분석을 한다. 지수평활 분석을 통해

미래의 핵심 키워드 출현 빈도를 예측하고 과거의 데이터와 비교하여 부상도를 계산한다. 부상도가 높은 핵심 키워드를 선정하여 부상기술을 예측한다.

4. 실험 및 결과

특허 데이터베이스 WIPS ON 을 통해 미국에서 출원된 탄소복합소재기술 특허를 2000 년부터 2009 년까지 출원연도별로 수집하였다.

<표 1> 연도별 출원 특허 수

출원연도	특허 문서	출원연도	특허 문서
2000	25	2005	63
2001	44	2006	86
2002	71	2007	71
2003	93	2008	69
2004	82	2009	82

수집된 특허 데이터 집합에서 발명의 명칭, 요약, 대표 청구항을 추출하였다. 추출된 데이터를 텍스트마이닝기법을 이용한 전처리 과정을 통해 분석이 가능한 문서-단어 행렬을 생성하였다. 문서-단어 행렬에 TF-IDF 가중치를 부여하고 30 개의 표본을 통해 정성적인 방법으로 도출된 최적의 TF-IDF 임계치 0.15 이상의 TF-IDF 가중치를 갖고 출현 빈도수가 200 개이상인 단어 21 개를 핵심 키워드로 선정하였다.

<표 2> 선정된 핵심 키워드

핵심 키워드	출현 빈도수	최대 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

핵심 키워드로 선정된 21 개 단어들의 2000 년부터 2009 년까지 연도별 출현빈도 수를 이용하여 지수평활 분석을 하였다. 이를 통해 2010 년의 핵심키워드의 출현빈도수를 예측하고 과거데이터의 평균과 비교하여 부상도가 높은 periodic, pinstock, microstructures, solution, leg, machine, outer 를 부상기술키워드로 선정하였다. 따라서 이 부상기술키워드와 관련이 있는 기술들의 연구개발이 필요하다.

<표 3> 지수평활을 이용한 부상도 계산

핵심 키워드	과거 데이터	예측데이터	부상도
impermeable	27.2	27.50729	0.30729
outline	27.5	26.85253	-0.64747
mat	26.5	27.1184	0.6184
ground	25.7	26.65923	0.95923
rolled	25.7	26.29012	0.59012
union	24.9	24.37397	-0.52603
sail-attaching	24	24.52106	0.52106
microporous	24.1	24.48903	0.38903
solution	23.8	25.06673	1.26673
microstructures	23.8	25.5762	1.7762
volumes	23.7	24.04764	0.34764
laminate	23.3	23.66961	0.36961
machine	23.1	24.16488	1.06488
permeable	23	22.89305	-0.10695
durable	20	19.21371	-0.78629
pumps	21.2	21.13003	-0.06997
leg	20.4	21.62617	1.22617
periodic	85.1	89.94971	4.84971
outer	45.9	46.90997	1.00997
responsive	43.8	42.48844	-1.31156
pinstock	37.4	39.87887	2.47887

5. 결론

본 논문은 기술예측을 위한 특허정보분석으로 텍스트마이닝 기법과 지수평활법을 이용한 방법을 사용하였다. TF-IDF 가중치를 이용하여 핵심키워드를 선정하고 지수평활법을 이용하여 부상도를 계산하였다. 결과적으로 periodic, pinstock, microstructures, solution, leg, machine, outer 를 부상기술키워드로 선정하였다. 향후 연구에서는 지수평활 외에 다른 시계열분석방법을 이용한 더욱 정확한 예측이 필요 할 것이다. 또한 선정된 키워드를 이용해 부상기술을 정의하는 방법에 대한 연구가 진행되어야 할 것이다.

감사의글

◆ 본 논문은 BK21 플러스 사업(고려대학교, 제조·물류분야에서의 빅데이터 운용 사업팀)으로 지원된 연구임.

◆ 본 논문은 2012 년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (한국연구재단-NRF-2010-0024163)

참고문헌

[1] 특허청 산업재산인력과 한국발명진흥회 산업인력양성팀. “특허와 정보분석(개정판)”, 경성문화사, 2009.
[2] Byungun Yoon, Yongtae Park.(2004). “A text-mining-based patent network: Analytical tool for high-technology trend”, Journal of High Technology Management Research, Vol.15(1), p.37-50.
[3] 전성해.(2011). “특허분석을 이용한 지능형시스템의 기술예측”, 제 21 권, 제 1 호, p.100-105.
[4] 김요섭, 박상성, 장동식.(2012). “CLARA 알고리즘을 사용한 특허정보 분석”, 제 10 권, 제 6 호, p.161-170.
[5] 최진호, 김희수, 임남규.(2011). “기술예측을 위한 특허 키워드 네트워크 분석”, 지능정보연구, 제 17 권, 제 4 호, p.227-240.
[6] Bowerman, O’Connell, Koehler. “Forecasting, Time series, and Regression(4th)”, BROOKS/COLE CENGAGE learning, 2005.
[7] WIPS ON. http://www.wipson.com, (March 12,2014).
[8] KIPRIS. http://kipris.or.kr, (March 12, 2014)