특허분석을 위한 군집화 알고리즘에 관한 연구

A Study on Clustering algorithm for Patent analysis

김 종 찬, 전 성 해*, 김 갑 조, 박 상 성, 장 동 식
고려대학교, 청주대학교*

Kim jong-chan, Jun sung-hae*, Kim kab-jo,
Park sang-sung, Jang dong-sik
Korea Univ., Cheongju Univ.*

요약

최근 특허전략의 중요성이 강조됨에 따라 국가와 기업들은 기술개발 및 기업경영정책을 수립하는데 특허정보를 활용하고 있다. 기술 경쟁력을 확보하기 위해서 특허정보분석을 통해 미래기술예측을 할 수 있다. 이 기술예측 프로세스에는 다양한 알고리즘들을 사용 된다. 본 논문은 특허분석을 이용한 기술예측 프로세스에 사용되는 군집화 알고리즘들에 대해 알아보고 각각의 특징과 장단점을 비교하였다. 이를 통해 분석대상과 방법에 따른 기존 알고리즘의 사용과 특허 분석에 새로운 알고리즘 적용 방향에 대해 알아보고자 한다.

1. 서론

최근 특허전략의 중요성이 강조됨에 따라, 국가와 대부분의 기업들이 R&D 계획수립, 기술마케팅 등에 특허정보를 활용 하고 있다. 축적된 특허 정보를 분석함으로써 대상 기술 분야의 발전과정과 현재의 기술개발상황을 알아보고 향후의 기술개발 방향을 정립할 수 있다. 즉, 특허정보분석은 국가의 기술개발정책을 수립하거나 기업의 경영정책을 수립하는데 있어서의 R&D 중복투자를 방지하고 기술 경쟁력을 확보하기 위해 연구해야 할 공백기술 및 부상기술 분야를 추출하는 등에 목적이 있다[1].

본 논문은 공백기술 및 부상기술을 예측하기 위한 기술예측 프로세스에 사용된 기존의 군집화 알고리즘들을 알아보고 각각의 특징과 장단점을 소개하였다. 이를 통해 특허분석을 이용한 기술예측에 대해 체계적이고 효율적인 연구가 가능하도록 하고 기존의 알고리즘들을 개선하여 보다 효과적인 새로운 알고리즘을 개발하는데 도움이 되고자 한다.

2. 특허 분석

기술예측을 위해 군집화를 이용하는 특허 분석의 프로세스는 다음과 같다.

기술예측이 필요한 관심 기술 분야의 특허 데이터를 KIPRIS, WIPSON, USPTO 등과 같은 특허DB에서 수집한다. 수집 된 특허 데이터는 정량분석이 가능하도록 텍스트 마이닝을 이용한 전처리 과정을 거쳐 특허-키워드 행렬구조로 변환한다[2]. 이 특허-키워드 행렬로부터 주성분 분석을 통해 군집화에 이용할 주성분을 추출한다.

▶▶ 그림 1. 특허 분석 프로세스

그런데 특허-키워드 행렬은 관측치에 해당되는 특허보다 변수에 해당되는 키워드의 크기가 크므로 일반적인 주성분 분석을 사용하지 못하고 SVD(Singular Value Decomposition)을 이용한 주성분 분석을 사용한다[3]. 추출된 상위 주성분을 이용해 군집화를 한 후 그 결과를 해석하여 부상기술 및 공백기술을 예측한다.

3. 특허 군집화

데이터 마이닝 기법중 하나인 군집화 기법은 개체들 간의 유사성 또는 거리를 이용하여 비슷한 개체들끼리 집단화 하는 분석기법이다.

특허분석에서 군집화의 목적은 특허문헌에서 추출한 키워드들을 비슷한 특성을 가진 군집으로 집단화하여 각 군집을 통해 부상기술군 또는 공백기술군을 추출하는데 있다. 본 논문에서는 이러한 목적을 가지고 특허분석에 이용된 특허 군집화 알고리즘들을 알아보려고 한다.

4. 특허 군집화 알고리즘 비교

알고리즘	이용한 논문	장점	단점	특징
K-means	Using patent data for technology forecasting : china RFID patent analysis	사용이 쉽고 간편, 사전 정보 없이 의미 있는 결과 도출, 거의 모든 형태의 데이터에 적용 가능.	임의로 정하는 군집 수 K가 부적합한 경우 결과가 부정확, 비유사성 거리 정의와 가중치 결정이 어려움, 이상치 데이터에 민감함.	주어진 K개의 군집으로 집단화 할 때 각 군집과의 거리 차이의 분산을 최소가 되도록 함.
Agglomerative	Technology clustering based on evolutionary patterns: The case of information and communications technologies	군집화 과정이 트리 구조를 가지고 있어 어느 단계에서 군집화가 잘되었는지 확인이 용이함, 데이터 처리속도가 빠름,	한 번 군집화된 데이터는 수정되지 않음, 이상치 데이터에 민감함.	각 객체를 각각의 군집으로 배치하고 Euclidean 거리와 같은 측도들을 기준으로 가장 가까운 군집끼리 병합하는 방법.
SOM	DIVA: a visualization system for exploring document data bases for technology forecasting	개체의 위치를 시각적으로 보여줌, 빠른 처리 속도와 대용량 데이터 처리에 적합함.	군집 경계가 불명확함, 너무 많은 군집 수를 형성.	다수의 다차원 공간상의 개체들이 스스로 비슷한 것들을 찾아 2차원 공간에 자리 잡도록 하는 방법.

▶▶ 그림 2. 특허 군집화 알고리즘 비교

K-means 알고리즘은 사용이 간편하고 거의 모든 형태의 데이터에 적용 가능하여 특허분석에 가장 많이 사용되고 있는 군집화 알고리즘이다. 그러나 군집 수 K를 임의로 정해야하고 평균을 이용하기 때문에 이상치 데이터에 민감하다. Agglomerative 군집 알고리즘은 계층적 군집방법의 한 가지로 각 객체를 각각의 군집으로 배치하고 Euclidean거리 등을 이용해 가장 가까운 군집끼리 병합하는 방법이다. 군집화 과정이 트리 구조를 가지고 있어 어느 단계에서 군집이 잘되었는지 확인하기가 용이하고 데이터 처리 속도가 빠르다는 장점이 있다. 그러나 한 번 군집화 된 데이터가 수정 되지 못하고 K-means 알고리즘과 같이 이상치 데이터에 민감하다는 단점이 있다. 자기조직화지도(Self-Organizing Map)는 대용량의 데이터를 처리하는데 용이하면서도 빠른 처리속도를 가지고 있고 개체들의 위치를 시각적으로 보여 준다. 또한 입력벡터를 정해주면 K-means 알고리즘과 달리 군집 수를 정해주지 않아도 되지만 군집경계가 불명확하고 일반적으로 너무 많은 군집수를 형성한다는 단점이 있다 [4],[5],[6],[7].

5. 결론

군집화 알고리즘들은 서로 다른 특징과 장단점들을 가지고 있다. 기존 논문에 가장 많이 사용되고 있는 K-means 알고리즘뿐만 아니라 다양한 군집화 알고리즘들을 각각의 특징과 연구 목적에 맞게 사용함으로써 보다 좋은 결과를 기대할 수 있다. 또 군집 수를 임의로 정해야 하는 군집화 알고리즘을 사용할 때 임의성을 배제하기 위해 Davies-Bouldin Index, BIC, AIC, 실루엣과 같은 측도들을 사용한다. 이러한 최적군집 수를 구하기 위한 전처리 방법들에 대한 연구와 자기조직화지도 알고리즘의 단점인 너무 많은 군집 수가 생성되는 점을 보완하기 위해 K-means 기법을 사용하여 군집수를 줄여주는 것과 같이 군집화 알고리즘간의 상호보완 할 수 있는 방법이나 Raw 데이터의 필터링을 통해 노이즈를 제거하는 알고리즘 개발에 대한 연구는 향후 과제이다.

감사의 글

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

▌참 고 문 헌▐

[1] 특허청 산업재산인력과, 한국발명진흥회 산업인력양성팀, 특허와 정보분석(개정판), 경성문화사, 2009.

[2] Fattori, M., G., Turra, R.. "Text mining applied to patent mapping: a practical business case", World Patent Information, vol. 25, iss. 5, pp. 1-54, 2008.

[3] Hair, J. F., Black, B., Babin, B., Anderson, R. E., "Multivariate Data Analysis", Prentice Hall, 1992.

[4] Steven Morris et al, "DIVA: a visualization system for exploring document data bases for technology forecasting", Computers & Industrial Engineering, vol. 43, iss. 4, pp. 841-862, 2002.

[5] Hyoung-joo Lee, Sungjoo Lee, Byungun Yoon, Technology clustering based on evolutionary patterns: The case of information and communications technologies, Technological Forecasting & Social change, vol. 78, iss. 6, pp. 953-967, 2011

[6] Charles V. Trappey et al., "Using patent data for technology for technology forecasting: china RFID patent analysis", Advanced Engineering Informatics, vol. 25, iss. 1, pp. 53-64, 2011.

[7] 김현욱, 손철, 한상옥, 자기조직화지도를 활용한 동일강 수지역 최적 군집수 분석, 한국공간정보학회, vol. 20, no. 6, pp. 13-21, 2012.