특허분석을 위한 텍스트마이닝 기반 군집분석

Text-mining based on clustering for Patent Analysis

저자
(Authors)

이재봉, 정재우, 김응모
Lee Jae-bong, Chung Jae-woo, Kim Ung-mo

출처
(Source)

Proceedings of KIIT Summer Conference , 2017.12, 434-435 (2 pages)

발행처
(Publisher)

한국정보기술학회
Korean Institute of Information Technology

URL

http://www.dbpia.co.kr/Article/NODE07274937

APA Style

이재봉, 정재우, 김응모 (2017). 특허분석을 위한 텍스트마이닝 기반 군집분석. Proceedings of KIIT Summer Conference, 434-435.

특허분석을 위한 텍스트마이닝 기반 군집분석

이재봉(*), 정재우(**), 김응모(***)
(*) 성균관대학교 소프트웨어대학, vixlee2@skku.edu
(**) 성균관대학교 정보통신대학, cjw0828@skku.edu
(***) 성균관대학교 소프트웨어대학, ukim@skku.edu

Text-mining based on clustering for Patent Analysis

Lee Jae-bong(*) Chung Jae-woo(**), Kim Ung-mo(***)
(,***) Sungkyunkwan University, College of Software*
(**) Sungkyunkwan University, College of Information and Communication

요 약

최근 정보통신의 발달로 데이터의 양이 기하급수적으로 늘어나고 있고, 그 중 특허데이터의 양 또한 과거에 비해 매우 증가 되고 있는 추세이다. 따라서 특허 간의 관계가 복잡하게 얽혀 짐에 따라 특허정보 데이터베이스에 대해 정량적 또는 정성적 분석 기법 등 다양한 관점에서의 분석이 요구되고 있다. 본 논문에서는 이러한 특허데이터 특성을 파악하여 효율적으로 분석할 수 있도록, 효과적인 특허 데이터의 분석에 대한 방안으로 군집분석을 활용한 텍스트 마이닝 기법을 제시한다. 나아가 본 연구에서 제시한 방법이 유의미한 결과를 도출해 내는지에 대해 분석해 보고, 적용 가능성을 확인해 본다. 결론적으로 텍스트 마이닝 기반 특허 군집분석은 이질적 기술기반을 가지는 특허 간 비교에서는 유의미한 활용 가능성을 보여주었지만, 동질적 기술기반이 가지는 특허의 분석에서는 낮은 군집 성능을 보였다. 현재는 특허분석에 직접적으로 활용되기보다 보조적 수단으로 활용될 것으로 보이지만 분석기술이 발달하면 활용 될 여지가 있다.

1. 서론

현대 경영에 있어 특허분석은 기업이나 단체가 보유한 기술적 역량을 분석하는데 뿐만 아니라 특정 산업의 기술 동향을 파악하고 예측하는 데에도 활용되고 있다. 즉 특허를 분석하는 것이 기업의 기술전략 수립에 있어서 중요한 의사결정 도구로 활용되고 있다. 여타의 데이터 분석과 차별화 되는 특허 분석의 특징은, 발명이 가진 성질을 특정 하는 방법이 텍스트로 이루어져 있어 일반적으로 활용되는 데이터 분석 방식에 적합하지 않다는 것이다[1]. 따라서 특허분석은 주로 관련 전문가들이 직접 명세서를 읽고 분석하는 방식으로 이루어지며 자동화된 알고리즘이 개입될 여지가 크지 않다. 이러한 기존의 분석방식은 가속화되는 지식재산의 축적속도에 따라 분석에 소모되는 자원의 부담이 커지는 약점을 가지고 있다.

하지만 데이터 과학 영역의 발전에 힘입어 비정형 데이터를 분석하는 기술이 개발되면서 특허분석은 새로운 국면을 맞이하게 된다. 텍스트로 이루어진 비정형데이터를 키워드로 분해된 정형데이터로 만드는 텍스트 마이닝 기술은, 많은 연구자들에 의해 다양한 형태로 활용되게 된다. 특허분석의 영역 또한 이를 연구하는 연구자들에 의해 그 활용방법이 다양하게 제시되고 있다. 특히 텍스트 마이닝 기술은 다량의 데이터를 유사도에 따라 군집화 하는 군집분석과 함께 활용되면서 특허를 기술적 특성 또는 산업분야에 따라 분류하여 시각화 하는 특허맵 작성 부문에 적용하는 방법이 많이 논의되고 있다.

따라서 본 연구는 특허분석과정에서 자동화된 군집분석 알고리즘이 어느 정도 정확성을 띄는지 군집분석 방법을 통해 분석해 보고 현업에서 적용 가능성을 확인해본다.

2. 연구방법

2.1 와드방법(Ward Method)

계층적 군집분석인 와드 방법(Ward Method)은 매 단계마다 전체 군집 내의 제곱 합을 계산하여 이를 가장 최소화 하는 방향으로 새로운 군집 쌍을 형성하는 방법으로 군집의 개수를 줄여 나가는 방법이다.

와드 방법을 활용한 계층적 군집분석은 다음과 같은 절차를 따른다. 먼저 분석의 대상이 되는 각 객체 당 하나의 군집에 포함되어 있다고 간주한다. 이후 현재의 군집해에 있는 모든 군집 간의 쌍으로 묶었을 때 전체 제곱합을 산출한 후, 그중 전체 제곱합이 가장 작아지는 군집을 하나의 군집으로 묶는다. 이러한 과정을 모든 객체가 하나의 군집에 포함될 때까지 반복한다. 이런 군집화 과정은 덴드로그램을 통해 시각화 할 수 있다.

2.2 K-means 알고리즘

K-means 알고리즘은 비계층적 군집분석 방법 중 가장 널리 사용되는 군집분석이다. K개의 군집 중심좌표를 설정 한 후, 각 객체를 가까운 군집에 배정하는 알고리즘이다. 이 방법은 대량의 자료에서 군집을 발견하는데 효과적인 방법이다.

K-means는 다음과 같은 순서로 진행된다. 첫 번째는 사전 분류된 특허정보들을 기반으로 군집의 수를 설정하는 단계이다. 일반적으로 군집분석에서는 군집내의 분산과 관련한 통계량을 통해 최적의 군집 수를 추정하는 방법을 사용한다. 두 번째 단계는 텍스트마이닝 절차를 통해 도출된 단어-문서행렬(Term-Document Matrix)에서 임의의 k개의 중심좌표를 생성한다. 이 임의의 중심좌표는 개별 클러스터의 중심좌표로 설정되며, 추후 반복 알고리즘을 통해 이동하게 된다. 세 번째 단계에서는 설정된 중심좌표와 개체 간의 유사도를 계산하여 개체를 유사도가 가장 높은 군집에 배치시킨다. 네 번째 단계에서는 세 번째 단계에서 도출된 군집들 간의 중심벡터를 구한다. 이후 단계는 이렇게 새롭게 설정된 군집의 중심좌표와 이전 좌표 값을 비교하여 일정한 수렴조건에 드는 경우 종료되며, 그렇지 못한 경우 두 번째 단계를 반복하게 된다.

2.3 군집의 평가

군집분석은 사전에 개체의 속성이 분류되어 있지 않다는 특성상 외부적 척도를 통해 군집해의 유효성을 평가하기 위한 척도를 활용하기가 어렵다. 하지만 이미 사전 분류를 통해 모범적 군집해가 형성되어 있는 경우 분류분석과 유사한 방식으로 군집해의 성능을 평가 할 수 있고 이를 외부 평가지수라 한다. 외부평가지수로는 랜드지수(Rand Index)와 Purity가 흔히 사용된다.

3. 결론

분석에 사용한 군집분석 알고리즘을 평가한 결과, 이질적 기술기반을 가진 기업들을 대상으로 한 특허분석에서는 모든 군집 알고리즘과 유사성 척도가 유의미한 군집을 형성할 수 있었다. 하지만 동일한 기술기반 상 다른 사업 분류에 속한 특허들을 분석함에 있어서 군집분석 방법과 무관하게 유사성 척도로 유클리드 거리를 사용한 경우에 유의미한 군집을 얻을 수 없었고, 코사인 유사도를 사용하였을 경우 어느 정도 유의미한 군집해를 얻을 수 있다는 사실을 확인할 수 있었다.

결과적으로 텍스트 마이닝을 통한 군집분석은 그 알고리즘과 유사성의 척도의 활용에 따라 어느 정도 유의미한 결과를 도출할 수는 있으나, 실제 현업에서 활용하기에는 그 정확도가 상당히 떨어진다고 평가된다.

본 논문에서는 특허 분석의 방법을 군집분석으로 한정 지었지만, 기계학습을 통해 사전에 학습시킨 특허정보를 이용하는 분류분석 또한 활용가능성이 충분할 것으로 보인다. 또한 텍스트 마이닝의 기술영역이 단순한 키워드의 추출을 넘어서 키워드 간의 의미망(Semantic web)을 구축하는 단계를 지향하고 있고, 또한 문서 분석에 특화된 군집분석 방법이 개발되고 있는 만큼 정확도에 있어서 지속적인 개선이 이루어 질 것이라 기대된다.

참고 문헌

[1] 진성해, 엄대호, “특허와 통계학, 그 연결은?”, 한국통계학회논문집 제17권 제2호, 2010.

[2] T. Yuen-Hsien, “Text Mining for Patent Map Analysis”, World Patent Information vol 25, 2007.

[3] ALP Workshop on Patent Corpus Processing, 2003.

[4] T. Yuen-Hsien, “Text Mining Techniques for Patent Analysis”,Information Processing and Management 43, 2007.

[5] W. Agus, B. indra, “Clustering patent document in the field of ICT”, International Conference on Semantic Technology and Information Retrieval 28-29, 2011.

[6] J. Sunhae, P. Sangsung, J. Dongsik, “Technology forecasting using matrix map and patent clustering”, Industrial Management & Data Systems Vol.113 No.6, 2013.

[7] A. Gionis, H. Mannila, “Assessing data mining results via swap randomization”, ACM Transaction on Knowledge Discovery from Data vol.1 no.3, 2007.

[8] 전진, “텍스트마이닝 기반 특허 군집분석의 유효성 검증”, 학위논문, 2015.