Journal of the Korean Data & Information Science Society
2016, **27**(5), 1273–1284

http://dx.doi.org/10.7465/jkdi.2016.27.5.1273
한국데이터정보과학회지

키워드 네트워크의 클릭 분석을 이용한 특허 데이터 분석

김현희[1] · 김동건[2] · 조진남[3]

[123]동덕여자대학교 정보통계학과

접수 2016년 7월 11일, 수정 2016년 8월 8일, 게재확정 2016년 8월 30일

요 약

본 연구에서는 기계 학습 분야의 특허를 수집하여 키워드 네트워크를 구축하고 클릭 분석을 실시하였다. 먼저 텍스트 마이닝 기법을 적용하여 핵심 키워드들을 선정한 다음, 이 키워드를 기반으로 키워드 네트워크를 구축하였다. 다음으로 네트워크 구조 분석, 중요 키워드 분석 및 클릭 분석을 시행하여 2005년도와 2015년도에 출원된 기계 학습 특허의 동향을 파악하였을 뿐만 아니라 양해년도의 분석 결과를 통해 특허 경향을 파악하였다. 분석 결과 기계 학습 특허의 키워드 네트워크는 밀도와 군집계수가 낮은 것으로 드러났으며 기계 학습 기법 자체에 대한 특허보다는 다양한 응용 영역에서 기계학습을 적용한 특허들이 다수이기 때문으로 판단된다. 클릭 분석 결과 2005년도 클릭 분석에 의해 발견된 주제는 뉴스메이커 검증, 상품 소비 예측, 바이러스 공격 예방, 바이오마커, 그리고 워크플로우 관리였으며, 2015년도 기계 학습 특허 주제는 디지털 이미지 편집, 직불카드, 수신자 인라이닝 시스템, 유방 촬영 시스템, 재고 관리 시스템, 이미지 편집 시스템, 비행기 티켓 가격 예측, 그리고 문제 예측 시스템으로 나타났다. 2005년도에 비하여 2015년도의 근접 중앙성은 낮아지고 매개 중심성은 높아진 것으로 보아 최근의 특허 경향은 보다 다양한 분야에서 출원되고 있으며 이들 간의 연결이 활발해지고 있음을 알 수 있다. 클릭 분석은 클릭을 형성하는 키워드 집합을 해석하여 주제를 파악하는데 활용될 수 있을 뿐만 아니라 추출된 공유 멤버쉽 키워드 집합은 특허 검색 시스템과 같이 키워드 검색 기반의 시스템에서 검색 키워드로 활용될 수 있을 것으로 기대된다.

주요용어: 기계 학습 특허, 클릭 분석, 키워드 네트워크, 특허 분석.

1. 서론

과학 기술의 발전 속도가 빠르게 증가함에 따라서 현재의 기술 동향을 파악하고 미래의 기술 경향을 예측하는데 특허 정보의 분석이 필수적이다. 특허 문서에는 명시하고자 하는 기술의 배경이 된 지식, 해당 기술에 대한 상세한 설명뿐만이 아니라 산업상의 이용 가능성까지도 구체적으로 서술되어 있다. 따라서 특허 정보를 분석하면 특정 기술 분야의 핵심이 될 수 있는 원천 기술과 상대적으로 기술 개발이 취약한 공백 기술 및 기술의 발전 경향을 파악할 수 있다.

특허 문서는 자연어로 기술된 텍스트 문서이므로 이를 분석하기 위해서 텍스트 마이닝 기술이 요구되며, 트위터나 페이스북과 같은 SNS에서 파생된 문장들보다 정제된 언어로 기술되기 때문에 더욱 정확한 분석이 가능하다. 텍스트 마이닝 기법은 Lee 등 (2015)의 연구에서 온라인 영화 리뷰 분석에 적용되어 영화 흥행을 성공적으로 예측하는데 사용되었으며, Chae 등 (2013)의 연구에서 특허 문서 추천에 적

[1] (02748) 서울시 성북구 화랑로 13길 60, 동덕여자대학교 정보통계학과, 조교수.
[2] (02748) 서울시 성북구 화랑로 13길 60, 동덕여자대학교 정보통계학과, 교수.
[3] 교신저자: (02748) 서울시 성북구 화랑로 13길 60, 동덕여자대학교 정보통계학과, 교수.
E-mail: jinnam@dongduk.ac.kr

용된 바 있다. Tseng 등 (2007)이 제시한 텍스트 마이닝 기반 특허 분석은 특허 문서의 초록이나 전문을 추출하여 전처리한 후, 키워드 추출 알고리즘을 이용하여 중요도가 높은 키워드를 추출한다. 이후 추출된 키워드들을 클러스터링 하거나 분류하여 특허 정보에서 중요한 키워드들을 파악하는 것이 그 핵심 기술이다. 이 방법은 방대한 특허 정보의 내용을 핵심 키워드로 나타냄으로써 특허의 내용 및 동향을 예측할 수 있다는 장점이 있다. 반면에 특허들 간의 연관성 파악이 어려우므로 특허 간의 연결 구조를 이해하는데 제약점이 있다.

특허 간의 연결 구조를 이해하는 데 사용되는 방법은 Erdi 등 (2013)이 사용한 특허 네트워크 분석 방법이다. 특허 네트워크는 해당 특허가 인용한 다른 특허들을 기반으로 구성되므로, 특허가 인용한 이전 특허의 지식을 반영한다. 따라서 시간에 따른 지식 혹은 기술의 변화 과정을 알 수 있고 인용 관계를 분석하여 특허간의 연결 관계를 해석할 수 있다. 그러나 인용 관계는 특허의 내용이 직접적으로 반영되었다고 볼 수 없으므로, 특허의 내용을 바탕으로 특허 간의 관계를 파악하는 데 어려움이 있다. Kang 등 (2015)의 연구에서는 사회 네트워크 분석 결과를 통해 도출된 그룹 변수들을 텍스트 마이닝하여 배구 경기력 분석을 실시하였다. 이와 같이 사회 네트워크 분석과 텍스트 마이닝 기법을 함께 사용하면 텍스트 내용을 고려하여 연결 관계를 파악할 수 있다.

본 연구의 목적은 특허 문서의 키워드 네트워크를 구축하여 네트워크 구조 분석을 통해 특허를 이루는 키워드들의 상호 작용을 파악하고 중요한 키워드를 추출하여 출원된 특허 기술의 구체적인 분야를 찾는 것이다. 분석을 위해 최근 많은 관심을 받고 있는 기계 학습 관련 미국 특허를 특허 포털 사이트인 KIPRIS (`http://www.kipris.or.kr`)로부터 수집하였으며, 2005년도와 2015년도에 출원된 특허 전체를 각각 분석하여 해당년도에서 나타난 특허 동향을 이해할 뿐만 아니라 양해년도 특허 기술의 비교를 통하여 기계 학습 특허의 경향을 파악하고자 하였다.

먼저 특허 문서의 내용을 고려하면서 전체적인 연결 관계를 파악하고자 키워드 네트워크를 구축하고 키워드 네트워크 분석 및 클릭 분석을 시행하였다. 먼저, 키워드 네트워크를 구축하는 데 있어 성패를 좌우하는 것이 키워드 선정이라는 것에 주목하여 TF-IDF (Term Frequency - Inverse Document Frequency) 가중치를 기준으로 중요 키워드를 추출한 다음, 각 중요 키워드와 동시에 사용된 키워드를 연결 관계로 하여 키워드 네트워크를 구축하였다. 네트워크 구축 방식은 Kim등 (2016)의 논문에서 제안한 방식을 일반화한 것이다. 키워드 네트워크 분석은 네트워크의 전체적인 구조를 분석하기 위한 그래프 레벨 분석과 키워드 자체의 중심성을 분석하기 위한 노드 레벨 분석으로 나누어 실시하였다. 그래프 레벨 분석을 위해 네트워크 중앙성 (centralization), 밀도 (density) 및 군집 계수 (clustering coefficient)를 측정하였으며, 노드 레벨 분석을 위해 각 키워드의 연결정도 중심성 (degree centrality), 근접 중심성 (closeness centrality), 그리고 매개 중심성(betweenness centrality)을 측정하였다.

클릭은 세 개 이상의 노드로 구성된 최대 완전 서브그래프 (maximal complete subgraph)로서 클릭에 속하는 모든 노드가 서로 직접적으로 연결되어 있다 (Kwahk, 2014). 이러한 특성 때문에 클릭 분석은 일반적인 소셜 네트워크에서 강한 연결 정도를 갖는 커뮤니티를 찾는데 활용되어 왔다. 본 연구에서는 클릭 분석을 키워드 네트워크에 적용하여 강한 상관관계를 갖는 키워드 집합을 추출하는데 활용하였다. 즉, 본 연구에서 제안하는 키워드 네트워크의 클릭은 특허 문서 집합에서 함께 사용된 횟수가 많은 키워드들의 집합으로서 연관 키워드 집합을 의미한다. 또한 클릭에 속하는 노드가 다른 클릭에도 동시에 포함되는 공유 멤버쉽 노드를 추출하여 특허 주제를 파악하는데 활용하였다.

분석 결과 기계 학습 특허의 키워드 네트워크는 밀도와 군집 계수가 낮은 것으로 드러났으며 이는 기계 학습 기법 자체에 대한 특허보다는 다양한 응용 영역에 기계학습을 적용한 특허들이 다수이기 때문으로 판단된다. 또한 클릭 분석을 통해 추출된 키워드들 역시 같은 결과를 나타냈다. 2005년도 기계 학습 특허 문서에서 클릭 분석에 의해 발견된 주제는 뉴스메이커 검증, 상품 소비 예측, 바이러스 공격 예방, 바이오마커, 그리고 워크플로우 관리였으며, 2015년도 기계 학습 특허 주제는 디지털 이미지 편집, 직불

카드, 수신자 인라이닝 시스템, 유방 촬영 시스템, 재고 관리 시스템, 이미지 편집 시스템, 비행기 티켓 가격 예측, 그리고 문제 예측 시스템으로 대부분 기계 학습 응용 시스템으로 밝혀졌다. 2005년도에 비하여 2015년도의 근접 중앙성은 낮아지고 매개 중심성은 높아진 것으로 보아 최근의 특허 경향은 보다 다양한 분야에서 출원되고 있으며 이들 간의 연결이 활발해지고 있음을 알 수 있다.

클릭 분석은 주로 사람들의 관계망에서 파벌과 같은 결속이 강한 그룹을 찾는데 활용되어왔고 키워드 네트워크에 적용된 예는 드물다. 키워드 네트워크에서 클릭의 의미는 중요도가 높은 키워드들 중에서도 반드시 연결되어 등장하는 키워드 집합으로서 매우 전문화된 영역에서 그 활용도가 높을 것으로 기대된다.

본 논문은 다음과 같이 구성된다. 2절에서 특허 분석에 대한 관련 연구들을 살펴보고, 3절에서 특허 키워드 네트워크 분석 방법에 대해 자세히 설명한다. 4절에서 특허 분석 결과를 서술하고 마지막으로 5절에서 결론 및 향후 연구를 제시한다.

2. 특허 분석

과학 기술의 동향을 분석하고 예측하는데 특허 정보의 관리와 활용이 중요해짐에 따라서 특허 분석에 대한 연구가 관심을 받고 있다. 특허는 출원 혹은 등록될 때 IPC (International Patent Classification) 코드에 따라 분류되는데 이는 너무 광범위하여 실제로 특허 문서에 어떤 내용이 포함되는지 알기가 어렵다. 따라서 특허 문서의 내용을 파악하기 위해서 텍스트 마이닝을 적용해 볼 수 있다. 텍스트 마이닝은 특허 문서에서 중요한 키워드를 추출하므로 특허의 내용을 분석할 수 있으며, 이러한 측면에서 키워드 추출 기법이 특허 텍스트 마이닝에서 핵심적인 기술이라고 볼 수 있다.

Choi와 Hwang (2014)은 키워드 네트워크를 구축하고 커뮤니티 분석을 하여 발광 다이오드 분야와 무선 광대역 기술 분야의 연구 동향을 파악하고 기술 간의 상호 작용을 파악하였다. Kim 등 (2016)의 연구에서는 키워드 네트워크의 커뮤니티 분석 결과에 소셜 네트워크 분석을 재적용하여 영향력이 큰 커뮤니티를 찾아내고 이에 속하는 키워드들을 중심으로 사물 인터넷 특허 분야의 동향을 파악하였다. 본 연구는 키워드 네트워크에 클릭 분석을 적용하여 서로 연관성이 높은 키워드 집합을 추출함으로써 기계 학습 분야 특허의 시기별 특허 동향을 분석하였다.

특허 문서로부터 핵심 키워드를 추출하는 기술은 특허 검색이 키워드 검색을 기본으로 이루어지므로 검색어 선정에 활용될 수 있다. Noh 등 (2015)은 키워드의 빈도수, 분산, 그리고 TF-IDF 가중치를 이용하여 특허 문서로부터 키워드를 추출한 다음, k-means 클러스터링과 엔트로피 값을 계산하여 세 가지 방법의 성능을 비교하였다. 비교 결과 TF-IDF 가중치를 적용하여 키워드를 추출한 방법이 가장 성능이 좋음을 보여주었다. Noh 등 (2015)에서 사용된 키워드 추출 방법은 세 가지 모두 빈도수가 높은 키워드에 중요도가 있음을 기반으로 한 방법이지만, Li 등 (2009)은 빈도수가 상대적으로 낮은 키워드 중에서 중요한 키워드를 찾을 수 있는 알고리즘을 제안하였다.

본 연구에서는 클릭 분석을 활용하여 강한 연결성을 갖는 연관 키워드 집합을 추출하였다. 클릭 분석은 일반적인 소셜 네트워크에 적용되어 커뮤니티를 탐지하는데 주로 사용되었으므로 키워드 네트워크에 적용된 예는 드물다. Kargar와 An (2011)은 그래프 데이터에서 r-clique을 찾아내어 키워드 검색을 하는 방법을 제시하였다. 그래프에서 클릭을 찾는 방법과 달리 연관 규칙 마이닝을 적용하여 강하게 연결된 서브그래프를 찾아내는 연구가 진행되었다. Choubey 등 (2012)은 그래프 구조에 Apriori 알고리즘을 적용하여 동시에 연결된 노드 집합을 찾을 수 있음을 보여주었다. 그래프에 연관 규칙을 적용할 경우 노드의 빈도수만을 고려하므로 빈도수가 높은 노드들의 집합을 찾게 된다. 특허 문서의 경우 핵심 기술에 대한 키워드는 오히려 드물게 등장하므로 본 연구에서 제안한 클릭 분석을 활용하면 중요도가 높으면서 반드시 동시에 등장하는 키워드 집합을 찾을 수 있다는 잇점이 있다.

3. 연구 방법

3.1. 분석 자료

본 연구에서는 기계 학습 관련 미국 특허 정보를 수집하여 분석하였다. 기계 학습은 인공 지능의 한 분야로서 최근 그 관심이 폭발적으로 증가하였다. 따라서 과거 출원된 특허와 최근 출원된 특허의 동향을 비교하고자 2005년도와 2015년도에 각각 출원된 특허 전체를 수집하였다. 특허 문서는 발명의 명칭, 출원번호/출원일, 등록번호/등록일, 출원인, 발명자, 요약문 그리고 명세서로 구성된다. 발명의 구체적인 내용은 명세서에 기술되어 있으나 발명의 핵심 키워드들은 요약문에 나타나므로 본 연구에서는 요약문만을 처리하였다. "machine learning"을 검색 키워드로 사용하였으며 소프트웨어 분야의 특허 동향으로 한정하기 위하여 IPC 코드 G06과 G08로 제한하여 검색하였다. 2005년도에 출원된 소프트웨어 분야 기계 학습 특허는 560개 그리고 2015년도에 출원된 2,000개 특허를 수집하였다.

3.2. 키워드 네트워크 구축 및 분석

키워드 네트워크 구축 방법은 Algorithm 3.1과 같다. 먼저 각 특허 초록 당 등장한 키워드 중에서 명사만을 추출하여 문서-단어 간 행렬을 생성한다 (1). 이 중에서 중요한 키워드를 선별하기 위하여 각 단어마다 TF-IDF 가중치 값을 계산한다. TF-DF 가중치 산출 방식은 식 (3.1)과 같다 (2). 산출된 TF-IDF 가중치 값을 정렬하여 상위 10%에 해당하는 가중치 값을 선정한다 (3). (3)에서 선정된 TF-IDF 가중치 값보다 큰 값을 갖는 단어들을 선별한 다음 (4), 이 단어들과 동시에 사용된 단어들을 상관 계수를 기준으로 하여 재추출하였다 (5). 마지막으로 <선정된 키워드, 연관 키워드>로 표현하여 이를 기반으로 키워드 네트워크를 구축한다 (6).

Algorithm 3.1 Construction of Keyword Network

1. Generating Term-Document matrix
2. Caluculating TF-IDF weight for each term
3. Finding an appropriate score of TF-DF weight by cutting the scores off in the top of 10%
4. Selecting the terms with TF-IDF weight value greater than the specified score
5. Extracting associated terms of the selected terms using the correlation value
6. Let m be the selected terms and n be the associated terms, then a keyword network is constructed with $< m, n >$pairs

Manning 등 (2008)에 따르면 Algorithm 3.1에서 사용된 TF-IDF 가중치는 식 (3.1)과 같이 나타낼 수 있다. 문서 j에 속하는 키워드 i의 가중치 $w_{i,j}$는 tf값과 df값의 곱이다. 여기서 $tf_{i,j}$는 문서 j에 나타난 키워드 i의 빈도수이고, $df_{i,j}$는 키워드 i를 포함하는 문서의 개수의 역수의 로그값이다. 키워드의 빈도수만을 고려하면 가장 빈도수가 높은 키워드들은 대부분 문서 집합에서 가장 일반적으로 사용되는 키워드들인 경우가 많다. 이러한 키워드들을 제거하고 더욱 구체적인 키워드들을 추출하기 위해서 키워드가 문서 집합에 등장한 횟수의 역수를 곱하면 문서 집합 전체에 많이 등장한 키워드들은 가중치 값이 줄어든다.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_{i,j}}\right) \tag{3.1}$$

또한 2005년도와 2015년도에 나타난 키워드들의 가중치 값이 상이하므로, 양해년도에서 중요도가 높은 키워드를 동일하게 추출하기 위해서 TF-IDF 값의 상위 10%에 해당하는 값을 기준으로 그 값보다 큰 값을 갖는 단어들만을 선정하였다. 이렇게 선정된 키워드만 독립적으로 보면 서로 다른 특허 분야에서 어떤 의미로 사용되었는지 알기 어려우므로 선정된 키워드들과 동시에 사용된 키워드들을 추출하여

키워드간의 상호 관계를 파악하고자 하였다. 문서-단어 행렬에서 사용된 키워드들은 전처리 과정을 거쳐서 명사로만 한정하였으며, 이는 분석 결과로 나타난 키워드들을 기준으로 특정 기술 영역을 해석하고자 할 때 형용사나 부사와 같은 품사들은 구체적으로 기술 영역을 나타낼 수 없기 때문이다. 즉 동일한 형용사는 다양한 명사와 함께 쓰일수 있으나 이 형용사가 기술 영역을 해석하는데 직접적인 영향을 주지는 않으므로 명사만을 고려하여 분석하였다.

Table 3.1은 수집된 기계 학습 특허 데이터에 대해 각각 수집된 특허 초록문 수, 전처리 후 추출된 명사 단어 수, TF-IDF 상위 10% 값보다 큰값을 갖는 선정된 단어 수, 그리고 선정된 단어들과 상관 계수 0.4이상을 갖는 단어들을 포함한 키워드 네트워크의 총 노드수를 나타낸다. 먼저 2005년도에 출원된 특허의 초록은 총 560개로서 이 특허 초록문으로부터 2,470개의 단어가 추출되었다. 여기서 TF-IDF 가중치 값을 상위 10% 이상으로 지정하여 선정된 단어 수는 437개이다. 10% 보다 하위로 선정할 경우 일반적으로 사용되는 키워드들이 자주 등장하였으며 15%, 20%, 25%를 기준으로 키워드들을 선정하여 실험한 결과 가장 적절한 값으로 나타났다. 마지막으로 437개의 단어에 각각 상관계수 값이 0.4 이상인 단어들을 재추출하여 전체 네트워크를 구축하였으며, 이때 사용된 전체 단어는 1, 118개이다. 2015년도의 경우, 수집된 특허 초록문이 2,000개이며 문서-단어 행렬에서 선정된 명사는 4,644개이다. 여기에 TF-IDF 가중치 값의 상위 10%에 해당하는 값 이상을 갖는 단어들을 선정하면 1,143개이고 네트워크 구축에 사용된 전체 노드 수는 1,365개이다.

Table 3.1 Patent data on machine learning

Year	# of abstract	# of terms	# of selected terms	# of nodes
2005	560	2,470	437	1,118
2015	2,000	4,644	1,143	1,365

텍스트 마이닝을 위해서 사용된 언어는 R이며, R에서 제공하는 tm 패키지와 sna 패키지를 활용하여 키워드 추출, 키워드 네트워크 구축 및 분석을 수행하였다. 양해년도의 특허 동향을 파악하기 위해서 키워드 네트워크 구조 분석, 키워드 분석, 그리고 클릭 분석을 실시하였다. 먼저 전체적인 키워드 네트워크의 구조를 비교하고자 양해년도 키워드 네트워크의 중앙성, 밀도 그리고 군집 계수 (Freeman, 1979)를 분석하였다. 다음으로 각 키워드 네트워크에서 연결 중심성, 매개 중심성, 그리고 근접 중심성이 높은 키워드를 분석하여 그래프 내에서 중요 역할을 담당하는 키워드 및 관련 기술을 파악하였다. 마지막으로 중요한 기술 분야를 파악하기 위해서 클릭 분석을 실시하여 서로 연관성이 높은 키워드 집합을 추출하고 이를 기반으로 기술 분야를 찾아내었다.

4. 분석 결과

4.1. 키워드 네트워크 구조 분석

Figure 4.1과 Figure 4.2는 각각 2005년도와 2015년도에 출원된 기계 학습 분야 특허 키워드 네트워크를 나타낸다. 그림에서 원의 크기는 키워드의 연결 중심성에 비례해서 표현되었다. 그림에서 보이는 바와 같이 연결이 되지 않고 분리된 그래프 집합이 많았고, 이러한 분리된 그래프들은 키워드 네트워크 구조 분석에 활용되지 않으므로 분석에서 제외하였다. 즉, 키워드 네트워크의 구조 분석을 위해서 생성된 네트워크의 최대 연결 그래프만을 활용하였다. 각 키워드 네트워크로부터 최대 연결 그래프를 추출한 네트워크는 Figure 4.3과 Figure 4.4에서 볼 수 있다. 마찬가지로 키워드의 크기는 연결 정도 중심성에 비례해서 표현하였다.

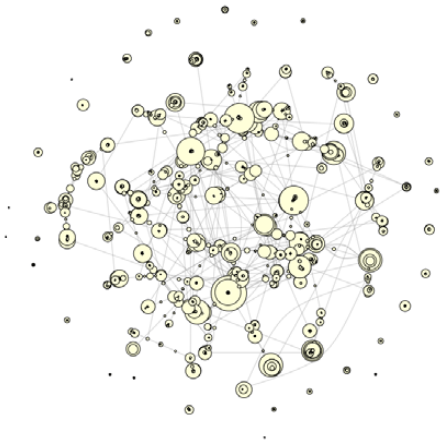

Figure 4.1 Keyword network in 2005

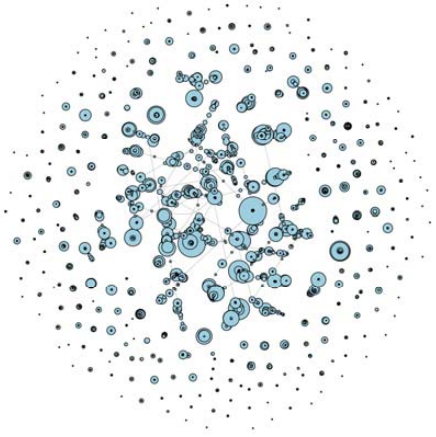

Figure 4.2 Keyword network in 2015

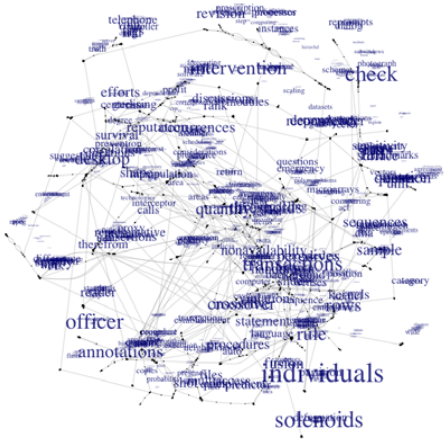

Figure 4.3 Maximal connected graph in 2005

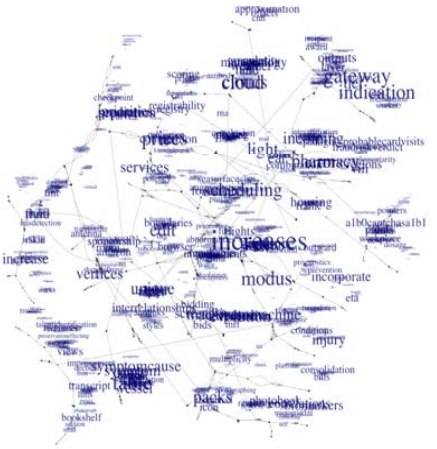

Figure 4.4 Maximal connected graph in 2015

키워드 네트워크의 구조를 파악하기 위해 키워드 네트워크의 중앙성, 밀도, 그리고 군집 계수를 각각 분석하였다. 중앙성이란 전체 네트워크의 형태가 얼마나 중앙에 집중되어 있는지를 나타내는 개념으로 네트워크의 전체적인 특성을 파악하기 위하여 사용되고 있다. 중심성을 측정하는 척도에 대해 각각 그 중앙성을 구할 수 있으며 Freeman (1974)에 의해 식 (4.1)과 같이 정의되었다.

$$C_A = \frac{\sum_{i=1}^{g} [C_A(n*) - C_A(n_j)]}{\max \sum_{i=1}^{g} [C_A(n*) - C_A(n_j)]} \tag{4.1}$$

$C_A(n_j)$를 한 노드의 중심성이라고 하고, $C_A(n*)$를 네트워크에 속하는 모든 노드들의 중심성 중에서 최대값이라고 하자. $\sum_{i=1}^{g}[C_A(n*) - C_A(n_j)]$는 최대 중심성값과 모든 노드의 중심성의 차들의 합이고, $\max \sum_{i=1}^{g}[C_A(n*) - C_A(n_j)]$은 논리적으로 가장 큰 차이를 합한 것이다. 중앙성은 연결정도 중심성, 근접 중심성, 그리고 매개 중심성에 대해 분석하였다.

네트워크 밀도는 가능한 연결선의 수 대비 실제 연결선의 수를 나타내는 값으로 네트워크에 속하는 노드들이 서로 얼마나 밀접하게 연결되어 있는지를 나타내는 척도이다. 네트워크 밀도가 큰 네트워크 일수록 각 노드들이 서로 활발하게 연결되어 있고 따라서 개개의 노드들이 중요한 역할을 담당한다고 볼 수 있다. 네트워크 군집 계수는 특정 노드와 이웃한 노드들이 서로 연결되어 있을 확률로서 군집 계수가 높을수록 노드들이 서로 잘 뭉치는 성질이 있다.

Table 4.1은 기계 학습 특허 키워드 네트워크의 구조 분석 결과이다. 전반적으로 양해년도에서 모두 네트워크 밀도와 군집계수가 매우 낮게 나타났다. 이는 기계 학습분야의 특허에서 사용된 키워드들이 서로 긴밀한 관계를 맺고 있지 않음을 나타낸다. 이는 4.3절의 클릭 분석의 결과로도 유추해 볼 수 있는데 기계 학습 분야의 특허가 대부분 기계 학습 자체보다는 다양한 응용 영역에 대한 특허들이므로 키워드간의 연결 정도는 낮은 것으로 해석할 수 있다.

Table 4.1 Keyword network structure analysis

Year	centralization			density	clustering coefficient
	degree	closeness	betweenness		
2005	0.01782	0.09501	0.13218	0.00246	0.24278
2015	0.01641	0.04556	0.46470	0.00265	0.26220

중앙성의 경우 2005년도에 비하여 2015년도의 매개 중심성이 큰 폭으로 증가한 것을 알 수 있으며, 반대로 근접 중심성은 큰폭으로 감소함을 알 수 있다. 2015년도 기계 학습 특허에서 근접 중앙성이 적게 나타난 것은 기계 학습 특허의 출원 분야가 2005년도에 비해 다양한 영역에서 나타난 것으로 파악된다. 반면에 매개 중심성의 값이 큰폭으로 상승한 것은 기술 분야 간의 매개 역할을 하는 키워드들이 다수 등장한 것으로 파악할 수 있다.

4.2. 중요 키워드 분석

키워드 네트워크에서 중요한 역할을 담당한 키워드들을 찾아보고 이를 통해 기계 학습 특허 동향을 파악하고자 하였다. Table 4.2와 Table 4.3은 각각 2005년도와 2015년도 기계 학습 특허 키워드 네트워크의 연결정도 중심성, 근접 중심성, 그리고 매개 중심성이 높은 상위 15개의 키워드를 정리한 것이다.

Table 4.2 Top 20 keywords in 2005

rank	keyword	degree centrality	keyword	closeness centrality	keyword	betweenness centrality
1	protein	0.01816	billing	0.00511	billing	0.11003
2	traffic	0.01541	length	0.00511	length	0.07742
3	contactor	0.01485	manager	0.00510	line	0.06524
4	billing	0.01431	introduction	0.00510	introduction	0.06252
5	communication	0.01376	extracts	0.00510	pathway	0.06224
6	pathway	0.01321	line	0.00510	scale	0.05801
7	matrix	0.01266	purposes	0.00510	seller	0.05730
8	consumption	0.01211	scale	0.00510	manager	0.05482
9	layer	0.01211	correctness	0.00510	traffic	0.05449
10	markers	0.01211	interval	0.00510	route	0.05384
11	reports	0.01158	production	0.00510	purposes	0.04785
12	cocktails	0.01158	deterioration	0.00510	attack	0.04772
13	vaccine	0.01158	pathway	0.00510	communication	0.04679
14	question	0.01101	commerce	0.00509	ddos	0.04641
15	grammar	0.01101	interests	0.00509	frames	0.04401

Table 4.3 Top 20 keywords in 2015

rank	keyword	degree centrality	keyword	closeness centrality	keyword	betweenness centrality
1	ticket	0.00862	ticket	0.00060	band	0.09835
2	waveguide	0.00862	protection	0.00060	airline	0.07737
3	passage	0.00729	price	0.00060	claim	0.07729
4	table	0.00663	flight	0.00060	operandi	0.07718
5	payment	0.00663	increases	0.00060	arm	0.07684
6	line	0.00630	uav	0.00060	thread	0.07615
7	ratio	0.00630	fuel	0.00060	vibrations	0.07528
8	student	0.00630	airline	0.00060	clientside	0.07525
9	microstructure	0.00630	ownership	0.00060	kidney	0.07469
10	plasma	0.00600	nozzle	0.00060	assembly	0.07453
11	mammograms	0.00563	claim	0.00060	injection	0.07327
12	ptnr	0.00563	datastore	0.00060	appliances	0.06901
13	radiation	0.00563	settlement	0.00060	separation	0.06890
14	assembly	0.00563	asking	0.00060	air	0.06880
15	course	0.00563	terminating	0.00060	retention	0.06812

먼저 연결 정도 중심성이 높은 키워드 집합은 다른 키워드들과 많은 연결을 가지고 있는 단어들로써 영향력이 큰 키워드 집합이라고 할 수 있다. 2005년도에 출원된 특허의 연결정도 중심성이 높은 키워드들은 protein, pathway, marker, vaccine등 생체 정보 분석에 사용되는 키워드들이 다수 포함되었으며, 2015년도에는 waveguide, plasma, mammogram, radiation등 유방 조영술에 관련된 키워드들이 다수 포함되었다. 근접 중심성이 높은 키워드는 네트워크에 포함된 다른 키워드들과의 거리가 가장 짧은 키워드로서 2005년도의 경우는 billing, production, deterioration, commerce, interests등 상품 수요 예측에 관련된 키워드가 주를 이루었고, 2015년도에는 flight, airline, ticket, price등 비행기 예약 시스템과 관련된 키워드가 주를 이루었다. 매개 중심성의 경우는 2005년도와 2015년도 모두에서 약 70%정도의 키워드가 겹쳐서 나타났으며, 이는 기계 학습 특허에서 근접 중심성이 높은 키워드들이 매개 중심성 역시 높게 나타났음을 알 수 있다.

4.3. 클릭 분석

양해년도 기계 학습 특허가 구체적으로 어떤 기술 분야의 특허인지 비교하기 위해서 클릭 분석을 수행하였다. 네트워크에서 클릭은 클릭 내의 모든 노드가 서로 직접 연결되고 네트워크 내의 다른 어떤 노드도 클릭 내의 노드와는 직접 연결 관계를 갖지 않는다는 특성을 갖는다. 즉 각 노드가 사람이라면 강한 연결 관계를 갖는 커뮤니티를 의미하고 각 노드가 상품이라면 강한 연결 관계를 갖는 상품 집합을 의미한다. 클릭이 키워드 네트워크에 적용되면 클릭에 의해 추출된 키워드 집합은 서로 직접적으로 연결된 최대 완전 서브그래프를 형성하므로 특허 집합 내에서 특정 기술을 나타내는 키워드 집합이라고 할 수 있다.

먼저 2005년도 키워드 네트워크의 클릭 분석 결과는 Table 4.4에 나타난다. 클릭의 크기는 서로 연결되어 있는 키워드 수에 따라서 3개부터 7개까지 나타났다. 이 중 최대 클릭을 이루는 클릭의 크기가 7인 키워드 집합과 그 다음 크기인 6인 키워드 집합을 분석하였다. 한 클릭에 소속된 키워드들이 다른 클릭에도 중복되어 소속될 수 있는데 특히 이러한 공유 멤버쉽에 속하는 키워드는 (comembership keywords) 전체 문서 집합에서 응집력이 강한 키워드들로서 다른 그룹과의 매개 역할을 수행할 수 있다. 클릭의 크기가 7인 클릭은 총 4개가 발견되었으며, 그 중 3개는 공유 멤버쉽 키워드를 갖고, 다른 1개는 다른 클릭과 연결되지 않은 독립적인 클릭으로 나타났다. 그룹 1은 공유 멤버쉽 키워드를 갖는 클릭으로 뉴스메이커 검증에 관한 키워드 집합이며, 공유 멤버쉽에 속하지 않는 키워드는 각각 names,

aggregator, 그리고 news이다. 그룹 2는 공유 멤버쉽을 갖지 않는 독립적으로 구성된 한 개의 클릭으로 상품 소비에 대한 시계열 예측에 관한 키워드 집합으로 나타났다.

Table 4.4 keyword sets of cliques in 2005

Group	Comembership keywords	Keywords not in comembership
1	article, comments, newsmaker, newsmakers, reader, verification	(1) names (2) aggregator (3) news
2	time, forecasting, commodity, individual, series, consumption, population	
3	consumption, amplitude, power, attacks, frequencies	(1) microprocessor (2) worms (3) viruses
4	markers, disorders, mood, panel, treatment	(1) compounds (2) diagnosis (3) kit (4) differentiation (5) therapies
5	workflow, dependence, scenario, history, usage, device	

클릭의 크기가 6인 클릭은 총 9개로서 그룹 3부터 그룹 5가 여기에 속하며 2개의 공유 멤버쉽을 갖는 클릭과 1개의 독립적인 클릭으로 나타났다. 그룹 3은 바이러스 공격 예방에 관한 공유 멤버쉽 키워드를 갖는 클릭으로 공유 멤버쉽에 속하지 않는 키워드는 microprocessor, worms, 그리고 viruses였다. 그룹 4는 바이오마커에 관련된 클릭으로 compounds, diagnosis, kit, differentiation, 그리고 therapies가 공유 멤버쉽에 속하지 않는 키워드로 나타났다. 마지막으로 그룹 5는 워크플로우 관리 등에 관련된 키워드 집합이 추출되었다.

2015년도 키워드 네트워크의 최대 클릭은 클릭 크기가 5인 클릭으로 총 27개였다. Table 4.5는 2015년도 클릭 크기가 5인 클릭의 키워드 집합을 나타낸다. 먼저 그룹 1에서 그룹 7까지는 공유 멤버쉽 키워드를 기준으로 분류한 것이며, 그룹 8은 독립적인 클릭으로 나타났다. 그룹 1은 디지털 이미지 편집, 그룹 2는 직불카드, 그룹 3은 수신자 인라이닝 시스템, 그리고 그룹 4는 유방 촬영 시스템에서 사용되는 키워드 집합이 공유 멤버쉽 키워드로 나타났다. 그룹 5는 이미지 편집 시스템, 그룹 6은 비행기 티켓 가격 예측, 그리고 그룹 7은 결점 진단 시스템에서 사용되는 키워드가 공유 멤버쉽 키워드로 나타났다. 독립적인 클릭인 그룹 8은 온라인 구매 시스템에 대한 키워드로 나타났다.

Table 4.5 Keyword sets of cliques in 2015

group	Comembership keywords	Keywords not in comembership
1	complexity, cropping, croppings, simplicity	(1) measuring (2) gradient (3) entropy (4) boundary
2	deposits, cash, deposit, brand	(1) authorization (2) payment(3) retailer
3	ratio, callee, costeffectiveness, inline	(1) reachable (2) ration(3) inlining
4	mammograms radiation, ptrn, dose	(1) xray (2) patch (3) tnr (4) microcalcifications (5) masses (6) higherdose
5	photos, spread, photobook, spreads	(1) photos (2) uploads (3) friction
6	price, protection, ticket, flight	(1) increases (2) terminating (3) asking (4) settlement
7	symptom, symptomcause, cause, fields	(1) table (2) names (3) associations
8	online, demand, timeexpiring, listing, inventory	

클릭 분석을 통해 발견된 키워드는 Kim등의 연구 (2016)에서 사용한 키워드 네트워크의 커뮤니티 분석 결과에 의해 발견된 키워드보다 무척 구체적이고 세부적인 영역을 지칭하는 키워드로 나타났다. 특히 클릭으로 드러난 키워드 집합은 양해년도 모두에서 기계 학습 자체에 대한 키워드보다는 기계 학습의 응용 분야에 대한 키워드로 나타났으며 이는 출원되고 있는 기계□□학습 특허들이 주로 기계 학습이 다양한 응용 시스템에 적용되고 있음을 반영한다. 또한 2005년도의 클릭 분석 결과는 2000대 초반 새롭게 등장하기 시작한 기술 분야들로 나타났으며 2015년도의 클릭 분석 결과 역시 최근 관심을 받고 있는 기술 분야로 나타났다. 즉, 기계 학습은 새롭게 등장하거나 관심을 많이 받고 있는 분야에 효율적으로 응용되어 오고 있다는 것을 알 수 있다.

5. 결론 및 향후 연구

본 연구에서는 기계 학습 분야의 특허 경향을 알아보기 위해서 2005년도와 2015년도에 출원된 특허들을 수집하여 키워드 네트워크를 구축하고 분석을 시행하였다. 먼저 키워드 네트워크 구축을 위해 TF-IDF 가중치를 이용하여 중요 키워드를 선정하고 동시에 등장한 연관 키워드들을 추출하여 비방향 그래프를 생성하였다. 키워드 네트워크 분석은 전체적인 네트워크의 구조를 파악하기 위해 중앙성, 밀도 및 군집 계수를 비교하였으며, 중요 키워드를 알아보기 위하여 연결정도 중심성, 근접 중심성, 그리고 매개 중심성이 높은 상위 15개의 키워드를 조사하였다. 마지막으로 키워드 네트워크의 최대 완전 연결 그래프인 클릭을 찾아내어 강한 연결 관계를 갖는 키워드 집합을 확인하였다.

먼저 2005년도와 2015년도의 키워드 네트워크 구조는 연결정도 중앙성, 밀도, 군집 계수에 있어서는 큰 차이를 보이지 않았다. 단 2005년도에 비해서 2015년도의 근접 중앙성이 크게 감속하였고 매개 중앙성은 크게 증가함을 알 수 있었다. 이는 2005년도보다 2015년도의 기계 학습 특허의 주제가 보다 다양해졌으며, 서로 상호간의 연결이 긴밀한 것으로 해석할 수 있다. 양해년도 모두 밀도 및 군집 계수가 낮게 나타났는데 이는 기계 학습 특허에 사용된 키워드간의 연결 정도가 강하지 않음을 나타낸다. 이는 클릭 분석의 결과에서 보여지는 바와 같이 출원된 특허들이 기계 학습을 적용한 응용 시스템이었으므로 키워드간의 상호작용은 드문 것으로 해석된다. 2005년도 특허의 연결 정도 중심성이 높은 상위 키워드에는 생체 정보 분석에 사용되는 키워드들이 다수 등장하였으며, 2015년도 특허의 연결 정도 중심성이 높은 상위 키워드에는 유방 조영술에 관련된 키워드들이 다수 등장하였다.

클릭 분석을 통해 추출된 키워드들은 매우 구체적으로 특정 분야를 설명하는 키워드 집합으로서 다양한 기계 학습 응용 분야를 유용하게 찾아낼 수 있었다. 2005년도 클릭 분석에 의해 발견된 주제는 뉴스 메이커 검증, 상품 소비 예측, 바이러스 공격 예방, 바이오마커, 그리고 워크플로우 관리였으며, 2015년도 기계 학습 특허 주제는 디지털 이미지 편집, 직불카드, 수신자 인라이닝 시스템, 유방 촬영 시스템, 재고 관리 시스템, 이미지 편집 시스템, 비행기 티켓 가격 예측, 그리고 문제 예측 시스템으로 나타났다. 즉 특허가 출원된 시기에 새롭게 등장하거나 관심을 받았던 기술 영역에 기계 학습이 적용된 응용 시스템들이 출원되었음을 알 수 있다.

클릭 분석은 클릭을 형성하는 키워드 집합을 해석하여 주제를 파악하는데 활용될 수 있을 뿐만 아니라 추출된 공유 멤버쉽 키워드 집합은 특허 검색 시스템과 같이 키워드 검색 기반의 시스템에서 검색 키워드로 활용될 수 있을 것이다. 특히 구글 특허 검색 시스템이나 특허청 특허 검색 시스템에 클릭에 의해 발견된 키워드를 입력하면 해당 키워드를 모두 포함하는 특허 문서를 바로 찾아낼 수 있다. 이러한 클릭의 특성은 키워드 검색을 기반으로 하는 다양한 검색 시스템에서 연관 검색어 추천과 같은 방식으로 다양하게 활용될 수 있을 것으로 기대된다.

현재 상관 계수를 가중치로 한 가중치 네트워크를 분석하는 작업을 진행하고 있다. 현 연구에서는 지정된 상관 계수 이상을 갖는 키워드는 모두 동일하게 중요한 키워드로 보고 네트워크를 구축하고 분석하였으나, 상관 계수가 높은 키워드일수록 중요도가 높은 키워드로 보고 방향성 가중치 그래프를 구축하고 이에 적절한 분석 방법을 제시할 예정이다. 또한 본 연구에서는 단음절로 된 명사만을 고려하여 키워드 네트워크를 구축하였는데 명사와 명사로 구성된 이음절 이상의 키워드를 다루는 방안에 대해서도 연구를 진행할 예정이다.

References

Chae, M., Kang, M. and Kim. Y. (2013). Documents recommendation using large citation data. *Journal of the Korean Data & Information Science Society*, **24**, 999-1011.

Choi, J. and Hwang, Y. S. (2014). Patent keyword network analysis for improving technology development efficiency. *Technological Forecasting & Social Change*, **83**, 170-182.

Choubey, A., Patel, R. and Rana, J. L. (2012). Graph based new approach for frequent pattern mining, **4**, 221-235.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P. and Zalányi, L. (2013). Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics*, **95**, 225-242.

Faust, K. (2006). Comparing social networks: Size, density, and local structure. *Advances in Methodology and Statistics*, **3**, 185-216.

Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. *Social Networks*, **1**, 215-239.

Huh, M. H. (2014). Introduction to social network analysis using R, Free Academy, Seoul.

Kang, B., Huh, M. and Choi S. (2015). Performance analysis of volleyball games using the social network and text mining techniques. *Journal of the Korean Data & Information Science Society*, **26**, 619-630.

Kargar, M. and An, A. (2011). Keyword search in graphs: Finding r-cliques, *Proceedings of the VLDB Endowment*, **4**, 681-692.

Kwahk, K. Y. (2014). Social network analysis, Cheongram Publisher, Seoul.

Kim, D. H., Kim, H. H., Kim, D. and Jo, J. (2016). Social network analysis of keyword community network in IoT patent data. *Journal of Applied Statistics*, **29**, 719-728.

Lee, S., Cho, J., Kang, C. and Choi, S. (2015). Study on prediction for a film success using text mining. *Journal of the Korean Data & Information Science Society*, **26**, 1259-1269.

Lee, S., Yoon, B. and Park Y. (2009). An approach to discovering new technology opprtunities: Keyword-based patent map approaches. *Technovation*, **29**, 481-497.

Li, Y. R., Wang, L. H. and Hong, C. F. (2009). Extracting the significant-rare keywords for patent analysis. *Expert Systems with Applications*, **36**, 5200-5204.

Manning, Chr. D., Raghavan, P. and Schutze, H. (2008). Introduction to Information Retrieval, Cambridge University Press, New York.

Noh, H., Jo, Y. and Lee S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis, *Expert Systems with Applications*, **42**, 4348-4360.

Tseng, Y. H., Lin, C. J. and Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing and Management*, **43**, 1216-1247.

Wasserman, S. and Faust, K. (1994). Social network analysis: Methods and applications, Cambridge University Press, New York.

Journal of the Korean Data & Information Science Society
2016, **27**(5), 1273–1284

http://dx.doi.org/10.7465/jkdi.2016.27.5.1273
한국데이터정보과학회지

Patent data analysis using clique analysis in a keyword network

Hyon Hee Kim[1] · Donggeon Kim[2] · Jinnam Jo[3]

[123]Department of Statistics and information Science, Dongduk Women's University

Received 11 July 2016, revised 8 August 2016, accepted 30 August 2016

Abstract

In this paper, we analyzed the patents on machine learning using keyword network analysis and clique analysis. To construct a keyword network, important keywords were extracted based on the TF-IDF weight and their association, and network structure analysis and clique analysis was performed. Density and clustering coefficient of the patent keyword network are low, which shows that patent keywords on machine learning are weakly connected with each other. It is because the important patents on machine learning are mainly registered in the application system of machine learning rather thant machine learning techniques. Also, our results of clique analysis showed that the keywords found by cliques in 2005 patents are the subjects such as newsmaker verification, product forecasting, virus detection, biomarkers, and workflow management, while those in 2015 patents contain the subjects such as digital imaging, payment card, calling system, mammogram system, price prediction, etc. The clique analysis can be used not only for identifying specialized subjects, but also for search keywords in patent search systems.

Keywords: Clique analysis, keyword network, machine learning patent, patent analysis.

[1] Assistant professor, Department of Statistics and Information Science, Dongduk Women's University, Seoul 02748, Korea.
[2] Professor, Department of Statistics and Information Science, Dongduk Women's University, Seoul 02748, Korea.
[3] Corresponding author: Professor, Department of Statistics and Information Science, Dongduk Women's University, Seoul 02748, Korea. E-mail: jinnam@dongduk.ac.kr