지능정보연구 제17권 제4호 2011년 12월(pp.227~240)

기술예측을 위한 특허 키워드 네트워크 분석

최진호
세종대학교 경영학과
(jhchoi@sejong.ac.kr)

김희수
세종대학교 경영학과
(passion2080@gmail.com)

임남규
세종대학교 전자정보공학대학 컴퓨터공학과
(continueing@gmail.com)

특허의 중요성이 커짐에 따라 특허분석의 중요성 또한 점점 커지고 있다. 특허분석은 네트워크 기반 방법과 키워드 기반 방법으로 나눠지는데 네트워크 기반은 특허 내부에 존재하는 세부 기술정보에 대한 분석이 불가능하다는 단점이 있고 키워드 기반은 기술정보간의 상호관계를 규명하지 못한다는 단점이 있다. 기존에 제시된 네트워크 기반 특허 분석과 키워드 기반 분석의 한계를 극복하기 위해서 두 방법을 혼합한 방법으로서 본 연구에서는 특허 키워드 네트워크 기반 분석 방법론을 제시하였다. 본 연구에서는 LED 분야의 특허들을 대상으로 텍스트 마이닝을 통해 중요한 기술정보를 추출한 다음, 키워드 네트워크를 구축하고, 이를 대상으로 커뮤니티 네트워크 분석을 수행하였다. 분석 결과는 다음과 같다. 첫째, 특허 키워드 네트워크는 매우 낮은 밀도와 매우 높은 클러스터링 지수를 나타내었다. 밀도가 높다는 것은 LED 분야내 특허 키워드 네트워크 내 노드(키워드)들이 산발적으로 연결되어 있다는 것을 의미하며, 클러스터링 지수가 높다는 것은 해당 키워드 네트워크 내 노드, 즉 키워드들이 각각의 커뮤니티로 매우 긴밀하게 연결되어 있음을 나타낸다. 둘째, 특허 키워드 네트워크도 다른 지식네트워크와 마찬가지로 명확한 멱함수 분포를 따른다는 사실을 알 수 있었다. 이는 기존에 활발히 연구, 활용되어 많은 연결고리를 갖고 있는 특허개념(키워드)수록 지속적으로 다른 연구자들에 의해 선택되고 이 키워드를 바탕으로 새로운 키워드들이 연결되어서 이들 키워드간의 조합으로 새로운 기술이 발명된다는 것이다. 셋째, 특허가 개발될 때 특정 분야에 유입된 키워드 중 새로운 링크가 생긴 키워드의 대부분이 기존에 연결되어 있던 커뮤니티 내의 키워드들과 결합되어 새로운 특허 개념을 구성한다는 사실을 발견하였다. 이러한 사실은 단기(4년) 장기(10년) 두 기간 모두 동일하게 나타났다. 나아가 본 연구에서 제시한 방법론을 통해 도출된 특허 키워드 조합 정보를 활용하면 미래에 어떤 개념들이 합쳐져서 새로운 특허 단위로 만들어질지 가늠해볼 수 있고, 새로운 특허를 개발할 때 참고할 수 있는 유용한 정보로 활용할 수 있다.

논문접수일 : 2011년 11월 18일 게재확정일 : 2011년 12월 19일
투고유형 : 학술대회우수논문 교신저자 : 최진호

1. 서론

인류가 발전함에 따라 과학기술이 사회에 미치는 영향력은 큰 폭으로 커지고 있다. 그로 인해 오늘날 대부분의 조직은 상품/서비스 시장에서 큰 영향력을 발휘하는 핵심기술을 앞서 발굴하기 위해서 핵심역량을 집중하고 있다. 이렇게 기술의 중요성이 커져가고 있는 가운데 범세계적으로 쏟아져 나오는 기술정보를 체계적으로 분석해 새로운 기술을 예측하는 것은 시간과 비용이 상당히 많이 들며 만족할만한 결과를 얻기가 쉽지 않다. 이에 기업 및 연구기관은 새로운 기술과 혁신적인 아이

* 이 논문은 2011년도 세종대학교 교내연구비 지원에 의한 논문임.
아울러 본 연구는 2011년도 한국지능정보시스템학회 추계학술대회에서 Fast Track 게재후보작으로 추천된 논문임.

디어를 얻기 위해 다양한 기술예측 방법론을 고안해 왔다.

이러한 기술예측 방법론들은 분석 특성에 따라 크게 두 가지로 분류 할 수 있다. 첫 번째는 전문가들의 회의 및 의견조율을 통해 기술동향을 관찰하는 정성적인 분석 방법론으로서 관련 수목법(Relevance trees method)(Bengisu and Nekhili, 2006), 델파이기법(Weaver, 1971)등이 있다. 두 번째는 정량적인 분석 방법론으로써 트랜드 분석(Trend impact analysis)(Agami et al., 2008), 계량서지학(NARIN, 1994), 특허분석(Lee et al., 2008) 등이 있다. 정성적 방법론은 많은 기술전문 인력과 시간이 요구되고 도출될 결과가 참여하는 전문가들의 주관적인 성향에 따라 영향을 많이 받기 때문에 조직들은 기술예측을 위해 정량적 분석 방법론을 활용하거나 정성/정량적 두 방법을 부분적으로 조합해서 사용하는 추세이다(Lee et al., 2009). 정량적인 분석 방법론 중 특허분석은 특허가 기술의 원천 정보이면서 상업적 가치까지 지니고 있다는 점에 착안해 특허에서 추출한 데이터를 기반으로 기술예측을 하는 방법론이다(Campbell, 1983). 특허의 중요성이 갈수록 높아지고, 정보기술의 발전에 따라 특허 데이터베이스에 대한 접근성이 높아지고 있어서 특허분석의 실용성은 높이 평가 받고 있다(Lee et al., 2009).

초기 특허분석 방법론은 여러 기술분야에서 출원된 특허 수를 계산해 상호 기술분야 특허 수를 비교함으로써 중요한 기술분야를 파악했다(Wartburg, 2005). 이러한 방법론은 수치적 결과를 제공하기 때문에 중요 기술분야가 무엇인지는 파악할 수 있다. 하지만 출원된 특허 수만 안다는 한계점 때문에 관심 기술분야 내에 존재하는 핵심기술은 파악하지 못한다. 이러한 문제점을 극복하기 위해 최근 다수의 특허분석 방법론들이 제시되고 있다. 그 중 특허간의 관계를 네트워크 관점으로 분석하는 네트워크 기반 특허분석(Wartburg, 2005; Yoon and Park, 2004; Li et al., 2007)과 특허 기술내용을 분석하는 키워드 기반 특허분석이 있다(Lee et al., 2009; Yoon and Park, 2007). 네트워크 기반 특허분석은 특허간의 인용 관계를 기반으로 네트워크를 구성한 다음 어떤 특허가 특허의 중요도 및 특허간의 상호관계 분석을 하기 때문에 해당 기술분야의 흐름과 전반적 상황을 거시적 관점으로 볼 수 있다는 장점이 있다(Yoon and Park, 2004). 반면 키워드기반 특허분석은 특허내용에서 의미 있는 기술정보를 추출하여 형태소분석을 하는 방법론으로서, 특허에서 언급되는 중요 기술요소에 대한 구체적 정보를 파악할 수 있다는 장점이 있다(Yoon and Park, 2007).

하지만, 네트워크 기반 특허분석은 기술 분야에서 영향력을 미치는 특허를 파악할 수 있고, 특허들간의 상호관계를 파악할 수 있다는 장점에도 불구하고 특허단위로 분석을 수행하기 때문에 특허 내부에 존재하는 구체적인 세부 기술정보까지는 분석하지 못한다. 반면 키워드 기반 특허분석은 특허내용을 기반으로 분석을 수행하기 때문에 특허 내의 핵심기술 정보를 파악할 수는 있지만, 다른 특허에서 사용되는 기술정보들과의 상호관계를 규명하지는 못하는 한계점이 있다. 본 연구에서는 앞서 언급한 두 방법론의 한계점을 해결하기 위해 두 방법론을 통합한 기술분석 및 예측 방법론을 제시하고자 한다. 즉, 개별 특허를 대상으로 텍스트 마이닝을 통해 중요한 기술정보를 추출한 다음, 키워드 네트워크를 구축하고, 이를 대상으로 커뮤니티 네트워크 분석을 수행한다. 이를 통해 특허 각각이 지니는 핵심기술요소에 대한 구체적인 정보를 파악할 수 있고, 이들 간의 연관관계를 분석할 수 있으며, 나아가 해당분야 전문가들에게 미래

에 개발 가능한 구체적인 기술요소들의 조합을 제시할 수 있다.

본 논문은 아래와 같이 구성하였다. 제 2장에서는 본 연구와 관련된 선행 연구를 소개한다. 제 3장에서는 본 연구에 사용되는 데이터를 소개하고 텍스트 마이닝을 통해 각 특허별로 의미 있는 기술 키워드를 추출하는 과정 및 추출된 키워드를 기반으로 구축된 커뮤니티를 분석하는 과정을 순차적으로 제시한다. 제 4장에서는 분석결과를 제시하고, 이를 통해 본 연구 방법론의 타당성과 유효성을 검증한다. 마지막으로 제 5장에서는 커뮤니티 기반 네트워크 분석이 의미하는 바와 한계점 그리고 본 연구의 향후 발전 방향을 제시한다.

2. 문헌연구

본 장에서는 특허분석을 위해 사용된 기존의 분석 방법론과 본 연구에 활용된 방법론을 소개한다.

2.1 특허 분석

특허는 기술구현 내용, 기술 분류코드, 인용정보, 소유자 정보 등으로 구성된다. 이 구성요소들을 분석함으로써 기술변화 트렌드, 기술수준, 기술의 상업적 가치 등을 파악 할 수 있다. 그러므로 연구개발 또는 기술정책 및 기술전략을 담당하는 관련자들에게 있어서 특허분석은 중요한 정보를 제공한다(Yoon and Park, 2004; Wartburg et al., 2005). 특허를 분석하는 방법은 네트워크 기반 방법과 키워드 기반 방법으로 구분할 수 있다.

2.1.1 네트워크 기반 특허분석

네트워크 기반 특허분석은 특허의 인용정보, 유사도 같은 특징을 통해 네트워크를 구성하고 이를 네트워크 관점으로 분석한다(Wartburg et al., 2005). Yoon and Park(2004)은 특허간의 유클리디안 거리를 측정하여 네트워크를 구성한 다음 해당 기술분야에서 중요도가 높은 특허를 파악하였다. Wartburg et al.(2005)은 특허 인용 네트워크를 구성해 기술 클러스터 그룹을 파악하였다. 네트워크 기반 특허 분석은 특허들간의 상호연관성을 기반으로 분석함으로써 어떤 특허가 특정 기술 분야에서 영향력을 지니고 있는지에 대한 정보를 제공하고 기술분야의 거시적인 흐름을 파악 할 수 있다.

2.1.2 키워드 기반 특허분석

키워드 기반 특허분석은 각 특허가 담고 있는 문서 내용에서 핵심기술 정보를 파악하는 것을 말한다. 특허분석 분야에서 텍스트 마이닝이 사용되는 이유는 특허가 구조화 되지 않은 자연어로 구성되어 있어서 특허를 구조화된 데이터로 뽑아낼 필요가 있기 때문이다. 그 중 특허 데이터의 형태소 분석은 특허내용에 텍스트 마이닝을 수행하여 중요 기술 요소들을 포착하고 이를 형태학적으로 분석 하는 것을 말한다. 특허 데이터 형태소 분석의 한 예로서 Yoon and Park(2007)은 텍스트 마이닝을 통해 특허에서 키워드를 추출하고, 추출된 키워드를 벡터로 구성한 다음 형태소 행렬을 구성해 중요 기술에 대한 키워드 조합을 파악했다. 텍스트 마이닝을 통해 추출된 키워드 벡터들은 특허가 어떤 기술요소들로 구성 되어있는지 명시해 주기 때문에 기술 분야에서 활용되는 기술을 효과적으로 파악 할 수 있다.

2.2 네트워크 기반 커뮤니티 분석

네트워크 분석은 해당 요소간의 상관관계를 노드와 링크로 구성하고 노드들의 상호작용을 연구

하는 분야이다. 수학에서 파생된 이 분야는 현재 사회과학, 화학, 물리학 등 여러 분야에서 활용되고 있다. 사회과학분야의 예로는 주식시장의 네트워크 특성에 관한 연구(Vandewalle et al., 2001)와 언어 네트워크를 연구한 연구(Cancho and Richard, 2001)가 있다. 그 중 네트워크 기반 커뮤니티 구조 분석은 대상 네트워크를 하나의 전체집합으로 봤을 때 대상 네트워크 내에 존재하는 노드들을 대상으로 연결성이 강한 노드군을 파악하고, 이들 군들의 상관관계 및 특성을 분석하는 것을 말한다. 네트워크 안의 각 커뮤니티는 특정 이해관계 또는 배경을 가진 사회적 그룹으로 파악될 수 있다. 예를 들어 논문인용 네트워크 내 각 커뮤니티는 특정 주제에 관련된 논문들의 집합으로 볼 수 있다. 이를 통해 한 커뮤니티 안에 있는 노드, 즉 논문들은 같은 주제를 공유한다고 할 수 있다(Girvan and Newman, 2002). 이러한 네트워크 분석을 기반으로 연구주제 및 연구 트렌드에 대한 분석이 가능해지고 본 정보를 바탕으로 연구자는 자신의 연구 분야에 대한 정보를 효과적으로 파악할 수 있다.

3. 데이터 및 분석방법론

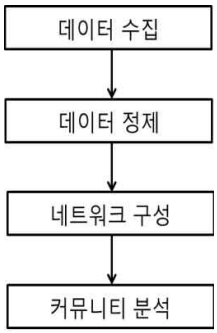
3.1 데이터

본 연구에서 사용된 분석 대상 자료는 LED(Light Emitting Diode) 분야의 특허로서, 미국 특허청(U.S. Patent and Trademark Office, USPTO)에서 출원된 국제 특허들을 대상으로 하였다. 추출된 특허는 WIPS[1]에서 검색어((light adj emi* adj diode) LED) and (fluorescen* adj material*)를 통해 검색된 2000년부터 2010년까지 출원된 LED 관련 314개의 국제 특허이다. LED 분야를 선택한 이유는 LED 분야가 최근 빠르게 발전해왔으며, 이로 인해 새로운 특허가 많이 생성되고 있으며, 특허 정보가 결합되는 과정을 효과적으로 볼 수 있기 때문이다.

1) http://www.wips.co.kr.

3.2 방법론

본 연구에서 제시하는 특허 키워드 네트워크 기반 기술예측은 아래 <그림 1>에서 제시된 바와 같이 네 단계로 나눠진다.



<그림 1> 본 연구의 분석 과정

3.2.1 데이터 수집 단계

첫 번째 단계는 데이터를 수집하는 과정으로서, 먼저 기술을 예측하고자 하는 특허분야를 정하고, 해당 분야 내 특허 데이터를 수집한다. 앞서 설명한 것처럼 본 연구에서는 특허 정보의 결합 과정을 효과적으로 파악하기 위해 새로운 특허가 많이 도출되는 LED 분야를 분석 대상으로 선정하고, 2000년부터 2010년까지 출원된 LED 분야의 특허 314개의 정보를 수집하였다.

3.2.2 데이터 정제 단계

두 번째는 자연어로 구성된 314개의 LED 분야 특허 초록에서 키워드를 추출하고, 이를 표준화 과

정을 통해 정제하는 단계이다. 특허 초록은 자연어로 구성되어 있기 때문에 먼저 텍스트 마이닝을 통해 특허 별로 중요한 키워드들을 추출한다. 다음으로 추출된 키워드 데이터들에 대해 해당 산업의 전문가의 도움을 받아 키워드들에 대한 표준화 작업을 수행하였다. 표준화 작업을 하는 이유는 텍스트 마이닝을 통해 추출된 키워드가 서로 동일한 뜻을 의미함에도 불구하고, 다른 단어로 쓰여져 있는 경우가 있기 때문이다. 표준화 규칙은 크게 아래의 세 가지 규칙을 기준으로 변경하였다.

- 복수형을 단수형으로 바꿈 : 예) DISPLAY SUBFIELDS → DISPLAY SUBFIELD
- 하이픈을 제거함 : 예) SECOND-HARMONIC LIGHT → SECOND HARMONIC LIGHT
- 동의어를 하나로 통일함 : 예) WHITE LED, WHITE LIGHT-EMIT DIODE → WHITE LED

3.2.3 네트워크 구성 단계

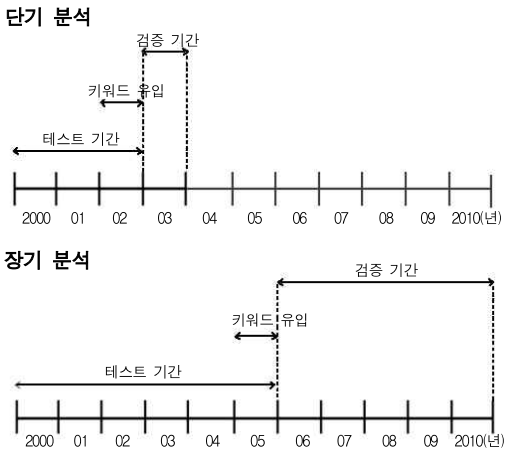
세 번째 단계로서 정제된 키워드들을 대상으로 네트워크를 구성한다. 앞서 언급한 것처럼 특허에서 추출된 각각의 키워드를 노드로 두고 키워드간의 관계를 링크로 맺는 네트워크를 구축하였다. 이렇게 네트워크를 형성한 이유는 텍스트 마이닝을 통해서 추출된 키워드는 특허에서 빈도수나 문맥에 따라서(Frantzi et al., 2000) 핵심이 되는 키워드이고 이 키워드들이 한 특허에서 도출되었다는 것은 이 키워드들이 핵심 특허 개념요소로 작용하여 하나의 특허, 즉 기술로 만들어 졌다는 것을 의미하기 때문이다. 이러한 네트워크 구성 방식에 기반하여, 지난 10년(2000~2010년)동안 출원된 LED 분야 특허들을 대상으로 가중치 네트워크(Bagler, 2007)를 만들었다.

3.2.3 네트워크 기반 커뮤니티 분석 단계

네 번째 단계는 구성된 네트워크를 이용해서 커뮤니티를 구성하고 분석하는 단계이다. 먼저 특허 키워드간의 상관관계를 기반으로 테스트 기간에 대한 키워드 커뮤니티를 구성하고, 새로 유입된 키워드들이 미래에 어떤 커뮤니티 내 키워드들과 링크를 하는지 확인한다. 이러한 확인을 통해 새로 유입된 키워드들이 기존에 관계를 맺고 있던 커뮤니티내 키워드들과 연결될 확률을 분석한다.

본 연구에서는 테스트 데이터 대비 예측 시점간의 기간을 기준으로 단기예측(3년치 데이터 기반 4년차 예측)과 장기예측(6년치 데이터 기반 6~10년차 예측) 두 경우로 나눠서 분석하였다. 두 경우에 대한 분석과정은 동일하다. 아울러, 실제 연결강도를 정확히 파악하기 위해 가중치 네트워크(Weighted network)를 기반으로 연결 정도를 분석하였다. 커뮤니티 분석은 Label Propagation 알고리즘(Raghavan et al., 2002)을 사용하였다.

단기분석은 <그림 2>와 같이 2000~2002년(3년) 구간을 테스트 기간으로 두고, 이 기간 대비 2002년에 새로 유입된 키워드가 테스트기간에 형성된 커뮤니티 내 키워드들에 연결되어 있는 비율을 분석하였다. 반면 장기분석은 <그림 2>에 제시된 것처럼, 2000~2005년(6년) 구간을 테스트기간으로 두고, 이 기간 대비 2005년에 새로 유입된 키워드들에 대해, 검증기간(2006~2010년)동안 다른 키워드들과의 모든 연결 수 대비 테스트 기간에 형성된 커뮤니티 내 키워드들과의 연결수의 비율을 분석하였다. 이를 통해 새로 유입된 키워드들이 기존에 관계를 맺고 있던 커뮤니티 내 키워드들과 연결될 확률을 파악하고자 하였다. 단기 및 장기 두 가지에 대해 분석한 이유는 본 연구에서 제시하는 분석 방법론이 시간적 간격에 상관없이 유의한 결과를 도출함을 보이기 위해서이다.



<그림 2> 연구 데이터 시간 구분

키워드 유입 기간을 1년으로 할당하였기 때문에 실험 네트워크를 구성하는 기간에 유입키워드 구성 기간을 포함 시켰다. 왜냐하면 두 기간을 단절하고 데이터를 수집 했을 경우 키워드의 유입 기간 1년 동안 실험 네트워크 구성 또한 변할 수 있기 때문이다. 그렇기 때문에 최신 네트워크 상태를 유지하기 위해 두 기간을 중첩하여 실험 데이터를 구성 하였다.

앞서 제시한 Label Propagation 알고리즘을 통해 테스트 기간에 출원된 특허들을 대상으로 추출된 키워드를 파악하고 이를 기반으로 네트워크 커뮤니티를 구성한다. 아울러 테스트 기간의 마지막 해에 유입된 키워드를 파악한다. 유입된 키워드들을 대상으로, 검증 기간 동안 새로 연결된 타 키워드들과의 모든 연결 대비 해당 키워드가 테스트기간동안 연결된 커뮤니티 내의 다른 키워드들과의 연결 정도를 분석함으로써, 새로운 연결이 기존에 형성된 커뮤니티 내에 많이 존재하는지 커뮤니티 외부에 많은 지 분석한다. 이를 통해 기술예측 및 분석에 있어서 신규 유입 키워드와 기존 커뮤니티 내 키워드들과의 조합을 통한 기술예측 가능성을 확인할 수 있다. 나아가 이러한 방법론에 기반하여 유입되는 키워드와 연결된 기존 커뮤니티에 대해 해당 키워드와, 연결된 커뮤니티 내 타 키워드들을 묶어서 제시함으로써 새로운 특허 및 기술개발을 위한 참조정보로 활용될 수 있다.

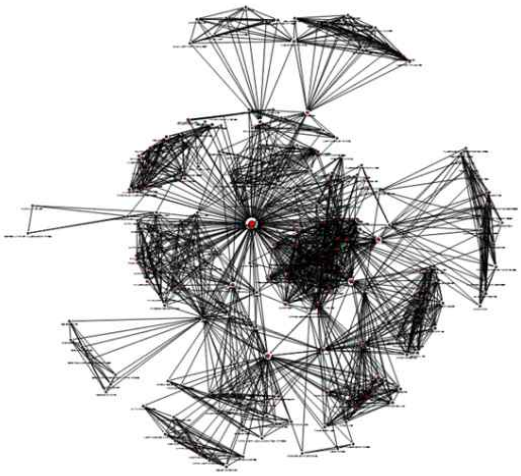
4. 분석결과

4.1 네트워크 주요 지표 분석

네트워크 기반 커뮤니티를 분석하기에 앞서 특허키워드 네트워크의 기본적인 특성을 파악하기 위해 먼저 주요 네트워크 지표에 대해서 분석하였다. 네트워크 밀도(Density)와 클러스터링 지수(Clustering coefficient)에 대한 분석 결과 <표 1>에서 제시된 바와 같이 특허 키워드 네트워크는 산발적이고 군집도가 높은 특성을 갖고 있는 것으로 파악되었다. 이러한 결과는 특허 키워드 네트워크의 한 예로서 <그림 3>에서 제시된 네트워크 구성도(2000~2003년 데이터 기반) 에서도 직관적으로 확인할 수 있다. 단기예측에 대한 분석결과 및 장기예측에 대한 분석결과 모두 공통적으로 매우 낮은 밀도와 매우 높은 클러스터링 지수를 나타내었다. 밀도가 높다는 것은 LED 분야내 특허 키워드 네트워크 내 노드(키워드)들이 산발적으로 연결되어 있다는 것을 의미하며, 클러스터링 지수가 높다는 것은 해당 키워드 네트워크 내 노드, 즉 키워드들이 각각의 커뮤니티로 매우 긴밀하게 연결되어 있음을 나타낸다. 즉 전체 특허 키워드 네트워크는 특정 주제별로 각각의 하부 커뮤니티로 나눠져 있고, 각 커뮤니티는 연관성이 높은 키워드들끼리 서로 강하게 연결되어 있으며, 각 커뮤니티 속에서 허브 역할을 하는 노드가 다른 커뮤니티와의 연결고리 역할을 하고 있음을 시사한다.

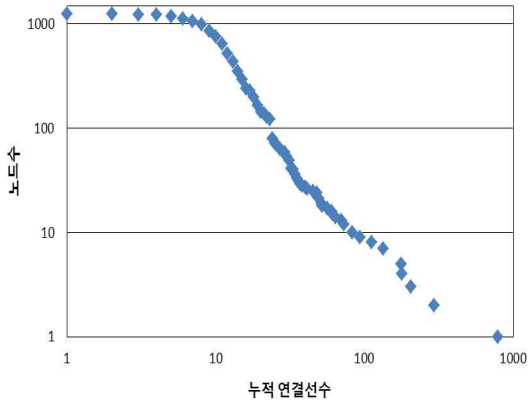
<표 1> 단기예측 및 장기예측 기간별 네트워크 지표

구 분	밀도	클러스터링 지수
단기(3년치 데이터 기반 4년차 예측)	0.011	0.924
장기(5년치 데이터 기반 6~10년차 예측)	0.034	0.946



<그림 3> 특허 키워드 네트워크 구성도 (2000~2003년)

아울러 특허 키워드 네트워크의 가장 흥미로운 특성이자 중요한 시사점 중의 하나로서, 네트워크의 누적 연결선 수 분포(cumulative degree distribution)(Albert et al., 2004)가 <그림 4>에서 제시된 것처럼 명확한 멱함수 분포(power-law distribution)를 따른다는 것이다. 멱함수 분포는 공동저자 네트워크, 인용 네트워크, 인터넷 네트워크 등(Jeong et al., 2008; Borner et al., 2004; Li et al., 2005; Wanga and Yua, 2008) 많은 네트워크 영역에서 관찰된다. 멱함수 분포의 가장 큰 특징은 부익부 빈익빈 현상(Buchanan, 2002)이다. 특허 키워드 네트워크에서 멱함수 분포가 의미하는 것은 기존에 활발히 연구, 활용되어 많은 연결고리를 갖고 있는 특허개념(키워드)수록 지속적으로 다른 연구자들에 의해 선택되고 이 키워드를 바탕으로 새로운 키워드들이 연결되어서 이들 키워드간의 조합으로 새로운 기술이 발명된다는 것이다. 특허 키워드 네트워크가 멱함수를 따른다는 사실은 본 연구를 통해 처음 밝힌 것으로서 본 연구의 가장 중요한 연구성과 중의 하나이다.



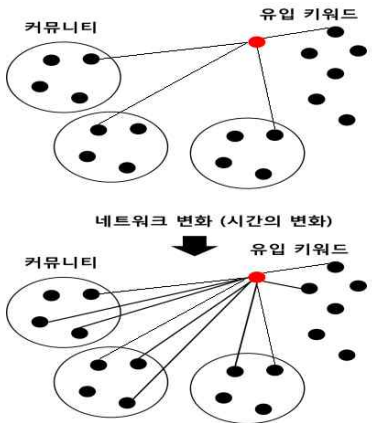
<그림 4> 누적 연결선 수 분포

4.2 커뮤니티 분석

<그림 5>에서 볼 수 있듯이 커뮤니티 분석의 단기예측 결과에 따르면 키워드 유입기간에 새로 유입된 키워드 중 새로운 링크가 생긴 키워드의 92%가 기존의 연결되어있던 커뮤니티 내의 키워드들과 링크가 생긴다는 사실을 확인할 수 있다. 장기예측 결과를 살펴보면 키워드 유입기간에 새로 유입된 키워드 중 검증 기간에 새로운 링크가 생긴 키워드의 86%가 기존에 연결되어 있던 커뮤니티 내의 키워드들과 새로운 연결을 한다는 것을 보여준다.

이러한 사실은 단기예측 및 장기예측 모든 경우에 있어서 새로 유입된 키워드에 대한 미래의 링크는 기존에 연결되어 있던 커뮤니티 내의 키워드

들과 연결될 확률이 매우 높다는 것을 보여준다.



<그림 5> 유입키워드의 링크과정

구체적으로 살펴보면, <표 2>에서 분석된 바와 같이 장기예측의 경우 키워드 유입기간에 새로 유입된 키워드 중 검증 기간에 새로운 링크를 맺은 횟수가 많은 상위 20개의 키워드 중 CONDUCTIVE MATERIAL만 빼고 나머지 14개 키워드는 모두 100% 기존에 연결되었던 커뮤니티 내의 키워드들과 연결되었다는 것을 알 수 있다. 이러한 사실은 대부분의 LED 관련 특허들은 특정 특허의 키워드가 기존에 링크되었던 커뮤니티 내의 다른 키워드들과의 조합을 통해 새로운 특허 단위로 개발된다는 것을 의미한다.

결론적으로, 본 연구에서 제시한 방법론을 통해 미래에 개발될 가능성이 높은 특허 구성요소(키워드)에 대한 정보를 세 단계를 거치면 쉽게 파악할 수 있다. 먼저 현시점에서 새로 출원된 특허에서 특허를 구성하는 주요 개념정보인 키워드들을 추출한다. 그 다음 커뮤니티 네트워크 분석을 통해서 새로 유입된 키워드가 기존의 어떤 커뮤니티와 링크를 하고 있는지 확인한다. 마지막으로 링크를 맺고 있는 커뮤니티 내의 키워드를 추출하여, 새로이 유입된 키워드와 조합한다. 이 세 단계를 통해 제시된 특허 키워드 조합 정보를 활용하면 미래에 어떤 개념들이 합쳐져서 새로운 특허 단위로 만들어 질지 가늠해볼 수 있고, 새로운 특허를 개발할 때 참고할 수 있는 유용한 정보로 활용할 수 있다.

4. 결론 및 향후 연구 방향

최근 특허의 중요성 및 특허 자료에 대한 접근성이 높아지고 있으며 특허분석의 실용성이 높이 평가 받고 있다.

이에 본 연구에서는 기존의 특허분석 방법론인 네트워크 기반 특허분석과 키워드기반 특허 분석의 한계점을 해결하기 위해 두 방법론을 통합한 기술분석 및 예측 방법론을 제시하였다. 특히 본 연구에서는 LED 분야의 특허들을 대상으로 텍스트 마이닝을 통해 중요한 기술정보를 추출한 다음, 키워드 네트워크를 구축하고, 이를 대상으로 커뮤니티 네트워크 분석을 수행하고, 이를 통해 특허 각각이 지니는 핵심기술요소에 대한 구체적인 정보를 파악할 수 있고, 해당분야 전문가들에게 미래에 개발 가능한 특허 키워드 조합을 제시하는 방법론을 제안하였다. 아울러 주요 네트워크 지표들을 중심으로 특허 키워드 네트워크가 갖고 있는 다양한 네트워크 특성을 분석하였다. 분석결과 특허 키워드 네트워크는 산발적이면서도 군집적으로 이뤄져 있으며, 네트워크 내 키워드 연결은 멱함수 분포를 나타내었다.

본 연구의 한계점으로는 LED 분야의 모든 특허를 수집하지 못한 점을 들 수 있다. LED 분야뿐 아니라 특정 분야의 특허는 그 양이 방대해서 특정 분야의 특허 데이터 추출 및 정제과정에서 많은 시간과 인력이 소요된다. 본 연구 대상 분야인 LED의 경우도 검색어를 LED로 할 경우 수만 개

<표 2> 커뮤니티 분석 결과

<단기 분석>

순위	키워드	기존커뮤니티 내 링크 확률	새로생긴 링크
1	LED CHIP	100.0%	18
2	LENS	100.0%	18
3	OPTICAL PROPERTY	100.0%	18
4	CONVEX FACE	100.0%	17
5	CUP PART	100.0%	17
6	SIMPLE PRODUCTION METHOD	100.0%	17
7	EXCESS RESIN MATERIAL	100.0%	17
8	RESIN MATERIAL	100.0%	14
9	TRANSPARENT LAYER SEAL	100.0%	8
10	REFLECTOR LAYER	100.0%	8
11	UPPER SIDE	100.0%	8
12	TRANSPARENT LAYER	100.0%	8
13	BLUE LED	16.7%	6
14	LIGHT EMITTING DEVICE	0.0%	5
15	FLUORESCENT LEAK DETECTION MATERIAL	100.0%	4
장기	전체 105개 키워드에 대한 평균 링크 예측치		새로생긴 평균 링크
	85.6%		10.72

<장기 분석>

순위	키워드	기존커뮤니티 내 링크 확률	새로생긴 링크
1	CONCAVE MIRROR	100.0%	40
2	TRANSPARENT REFLECTIVE OPTIC	100.0%	40
3	CURRENT REGULATOR CIRCUIT	100.0%	40
4	FLEXIBLE MEMBER	100.0%	40
5	POWER TYPE	100.0%	40
6	POWER LED	100.0%	40
7	ADDITIONAL OPTIC	100.0%	40
8	CONDUCTIVE MATERIAL	83.3%	36
9	MINIMUM SIZE HOLE PORTION	100.0%	26
10	INTERIOR SURFACE	100.0%	26
11	PROTECTIVE RESIN	100.0%	26
12	MAIN SURFACE	100.0%	24
13	IMAGING ELEMENT	100.0%	24
14	SEMICONDUCTOR ELEMENT	100.0%	24
15	IMAGING DEVICE	100.0%	24
장기	전체 105개 키워드에 대한 평균 링크 예측치		새로생긴 평균 링크
	85.6%		11.53

의 특허가 출원되기에 추출된 모든 특허들을 대상으로 하기에는 한계가 있었다. 이에 본 연구에서는 전문가의 자문을 얻어서 대상 분야를 좁혀서 분석하였다.

차후 연구주제로는 본 연구에서 제시한 방법론을 LED 이외의 다른 특허분야를 선택하여 해당 분야에서도 동일한 연구결과를 제시함으로써 본 연구방법론의 타당성을 검증한다. 나아가 두 분야 이상의 특허정보들이 결합하여 새로운 특허가 만들어지는 과정에서도 본 연구방법론의 적용 가능성 여부를 분석하고자 한다.

참고문헌

Agami, N. M. E., A. M. A. Omran, M. M. Saleh, and H. E. E. E. Shishiny, "An enhanced approach for Trend Impact Analysis", *Technological Forecasting and Social Change,* Vol.75, No.9(2008), 1439~1450.

Albert, R., I. Albert, and G. L. Nakarado, "Structural vulnerability of the North American power grid", *Physical Review E*, Vol.58, No.2 (2004).

Bagler, G., "Analysis of the airport network of India as a complex weighted network", *Physica A : Statistical Mechanics and its Applications,* Vol.387, No.12(2008), 2972~2980.

Bengisu, M. and R. Nekhili, "Forecasting emerging technologies with the aid of science and technology databases", *Technological Forecasting and Social Change,* Vol.73, No.7(2006), 835~844.

Börner, K., J. T. Maru, and R. L. Goldstone, "The simultaneous evolution of author and paper networks", *Proceedings of the National Academy of Sciences of the United State of America*, Vol.101, No.1(2004), 5266~5273.

Buchanan, M., "Nexus : small worlds and the groundbreaking theory of networks", *Local : Norton*, 2002.

Campbell, R. S., "Patent trends as a technological Forecasting Tool", *World Patent Information*, Vol.5, No.3(1983), 137~143.

Cancho, R. F. and V. S. Richard, "The small world of human language", *Proceedings of the royal society,* Vol.268, No.1482(2001), 2261~2265.

Frantzi, K., S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms", *International Journal of Digital Libraries,* Vol.3, No.2(2002), 117~132.

Girvan, M. and M. E. J. Newman, "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences of the United states of America,* Vol.99, No.12(2002), 7821~7826.

Jeong, H., Z. Néda, and A. L. Barabási, "Measuring preferential attachment for evolving networks", *Europhysics Letters,* Vol.61, No. 4(2003), 1~4.

Lee, S. J., B. G. Yoon, and Y. T. Park, "An approach to discovering new technology opportunities : Keyword-based patent map approach", *Technovation,* Vol.29, No.6/7(2009), 481~497.

Lee, S. J., S. H. Lee, H. J. Seol, and Y. T. Park, "Using patent information for designing new product and technology : keyword based technology roadmapping", *R&D Management,* Vol.38, No.2(2008), 169~188.

Li, M., Y. Fana, J. Chena, L. Gaoa, Z. Dia, and J. Wua, "Weighted networks of scientific communication : the measurement and topological role of weight", *Physica A : Statisti-*

cal Mechanics and its Applications, Vol.350, No.2/4(2005), 643~656.

Li, X., H. Chen, Z. Huang, and M. C. Roco, "Patent citation network in nanotechnology (1976~2004)", *Journal of Nanoparticle Research*, Vol.9(2007), 337~352.

Narin, F., "Patent bibliometrics", *Scientometrics*, Vol.30, No.1(1994), 147~155.

Raghavan, U. N., R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks", *Physical Review E,* Vol.76, No.5(2007).

Tseng, Y. H. and C. J. Lin, "Text mining techniques for patent analysis", *Information Processing and Management,* Vol.43, No.4(2007), 1216~1247.

Vandewalle, N., F. Brisbois, and X. Tordoir, "Non-random topology of stock markets", *Quantitative Finance,* Vol.1, No.3(2001), 372~374.

Wanga, M. and G. D. Yua, "Measuring the preferential attachment mechanism in citation networks", *Physica A : Statistical Mechanics and its Applications*, Vol.387, No.18(2008), 4692~4698.

Wartburg, I. V., T. Teichert, and K. Rost, "Inventive progress measured by multi-stage patent citation analysis", *Research Policy*, Vol.34, No.10(2005), 1591~1607.

Weaver, W. T., "The Delphi forecasting method," *The Phi Delta Kappan*, Vol.52, No.5 (1971), 267~271.

Yoon, B. G. 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(2004), 37~50.

Yoon, B. G. and Y. T. Park, "Development of new technology forecasting algorithm : hybrid approach for morphology analysis and conjoint analysis of patent information", *IEEE Transactions on Engineering Management,* Vol.54, No.3(2007), 588~599.

Yoon, B. U., C. B. Yoon, and Y. T. Park, "On the development and application of a self-organizing feature map-based patent map", *R&D Management,* Vol.32, No.4(2002), 291~300.

Abstract

Keyword Network Analysis for Technology Forecasting

Jinho Choi* · Heesu Kim* · Namgyu Im**

New concepts and ideas often result from extensive recombination of existing concepts or ideas. Both researchers and developers build on existing concepts and ideas in published papers or registered patents to develop new theories and technologies that in turn serve as a basis for further development. As the importance of patent increases, so does that of patent analysis. Patent analysis is largely divided into network-based and keyword-based analyses. The former lacks its ability to analyze information technology in details while the letter is unable to identify the relationship between such technologies. In order to overcome the limitations of network-based and keyword-based analyses, this study, which blends those two methods, suggests the keyword network based analysis methodology. In this study, we collected significant technology information in each patent that is related to Light Emitting Diode (LED) through text mining, built a keyword network, and then executed a community network analysis on the collected data. The results of analysis are as the following. First, the patent keyword network indicated very low density and exceptionally high clustering coefficient. Technically, density is obtained by dividing the number of ties in a network by the number of all possible ties. The value ranges between 0 and 1, with higher values indicating denser networks and lower values indicating sparser networks. In real-world networks, the density varies depending on the size of a network; increasing the size of a network generally leads to a decrease in the density. The clustering coefficient is a network-level measure that illustrates the tendency of nodes to cluster in densely interconnected modules. This measure is to show the small-world property in which a network can be highly clustered even though it has a small average distance between nodes in spite of the large number of nodes. Therefore, high density in patent keyword network means that nodes in the patent keyword network are connected sporadically, and high clustering coefficient shows that nodes in the network are closely connected one another. Second, the cumulative degree distribution of the patent keyword network, as any other knowledge network like citation network or collaboration network, followed a clear power-law distribution. A well-known mechanism of this pattern is the preferential attachment mechanism, whereby a node with more links is likely to attain further new links in the evolution of the corresponding network. Unlike general normal distributions, the power-law distribu-

* School of Business, Sejong University
** Department of Computer Engineering, Sejong univeristy

tion does not have a representative scale. This means that one cannot pick a representative or an average because there is always a considerable probability of finding much larger values. Networks with power-law distributions are therefore often referred to as scale-free networks. The presence of heavy-tailed scale-free distribution represents the fundamental signature of an emergent collective behavior of the actors who contribute to forming the network. In our context, the more frequently a patent keyword is used, the more often it is selected by researchers and is associated with other keywords or concepts to constitute and convey new patents or technologies. The evidence of power-law distribution implies that the preferential attachment mechanism suggests the origin of heavy-tailed distributions in a wide range of growing patent keyword network. Third, we found that among keywords that flew into a particular field, the vast majority of keywords with new links join existing keywords in the associated community in forming the concept of a new patent. This finding resulted in the same outcomes for both the short-term period (4-year) and long-term period (10-year) analyses. Furthermore, using the keyword combination information that was derived from the methodology suggested by our study enables one to forecast which concepts combine to form a new patent dimension and refer to those concepts when developing a new patent.

Key Words : Patent Analysis, Keyword Network, Text Mining, Technology Forecasting, Network Analysis

저 자 소 개

최진호

KAIST 산업경영학과에서 학사, 경영공학과에서 석사 및 박사학위를 취득하였으며, 현재 세종대학교 경영학과 조교수로 재직하고 있다. 주요 관심분야는 지식관리, 데이터마이닝 등이다. OMEGA, I&M, JASSS, Scientometrics 등의 국내외 학술지에 논문을 게재하였다.

김희수

현재 세종대학교 경영학과 학사과정에 재학 중이며 컴퓨터 공학을 복수 전공 중 이다. 주요 관심분야는 복잡계, 데이터마이닝, 네트워크 과학, 특허분석 등이다.

임남규

현재 세종대학교 컴퓨터공학과 학사과정으로 재학 중이다. 주요 관심분야로는 데이터마이닝, 분산처리, 소프트웨어 설계 등이다.