

Innovation of technology and social changes - quantitative analysis based on patent big data

Yongdai Kim^{a,1} · Sang Jo Jong^b · Woncheol Jang^a · Jongsu Lee^c

^aDepartment of Statistics, Seoul National University; ^bSchool of Law, Seoul National University;

^cTechnology Management, Economics and Policy Program, Seoul National University

(Received August 30, 2016; Revised October 10, 2016; Accepted October 10, 2016)

Abstract

We introduce various methods to investigate the relations between innovation of technology and social changes by analyzing more than 4 millions of patents registered at United States Patent and Trademark Office(USPTO) from year 1985 to 2015. First, we review the history of patent law and its relation with the quantitative changes of registered patents. Second, we investigate the differences of technical innovations of several countries by use of cluster analysis based on the numbers of registered patents at several technical sectors. Third, we introduce the PageRank algorithm to define important nodes in network type data and apply the PageRank algorithm to find important technical sectors based on citation information between registered patents. Finally, we explain how to use the canonical correlation analysis to study relationship between technical innovation and social changes.

Keywords: Patent, technical innovation, clustering, network analysis, canonical correlation

1. 서론

기술의 진보와 혁신은 사회의 발전을 추동하는 기본 동력이다. 기계에 의한 대량생산 시스템은 상업 자본을 대체하는 산업자본 시대를 열었으며, 최근에는 3D 프린터 기술의 발달로 거의 모든 사물을 프린트 할 수 있어 사실상 개개인 모두가 ‘생산수단’을 소유하게 되는 혁신적 사회 변화를 경험하고 있다. 또한 최근의 인터넷과 정보통신 기술의 발달로 인한 ICT융합기술의 급속한 발전은 사회 여러 분야에서 급격한 변화를 낳고 있다. 예를 들면, 다양한 의견이 인터넷을 통하여 여러 사람과 소통하면서 정치적 의사 결정 과정이 매우 빠르게 변하고 있으며, 사회적으로는 고위층 소수가 독점하던 정보의 한계가 없어지면서 다양한 계층들이 서로 융합하는 현상을 보여주고 있다.

이러한 시대적 배경에서 기술의 진보와 혁신이 사회의 변화에 미치는 영향을 과학적으로 분석하고 정리 하는 작업은 미래의 바람직한 사회건설을 위해서 매우 중요하다. 특히 최근의 급격한 새로운 기술의 등장과 이로부터 파생되는 빠른 사회구조의 변화는, 기술의 진보와 혁신이 사회변화에 미치는 영향을 거대 담론으로만 연구하는 것을 어렵게 만들었다. 이러한 상황에서, 기술의 진보와 혁신이 사회변화에 미치

This work was supported by SNU Brain Fusion Program of the Seoul National University in 2014.

¹Corresponding author: Department of Statistics, Seoul National University, 1 Gwanak-ro, Seoul 08826, Korea. E-mail: ydkim903@snu.ac.kr

는 영향을 정량적이고 객관적으로 분석, 정리 그리고 예측하는 연구가 절실히 요구되고 있다. 본 논문에서는 미국특허청에서 제공하는 데이터베이스를 사용하여 기술의 진보와 혁신이 사회변화에 미치는 영향을 정량적으로 측정하는 다양한 분석방법론에 대해서 소개하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 먼저 미국특허법의 변천사를 간략히 살펴본다. 특히, 특허법의 개정과 특허의 양적/질적 변화의 관계를 파악하여 특허의 양적/질적 변화 중에서 특허법의 개정으로 생긴 부분을 고찰하고자 한다. 3장에서는 국가별 특허 등록수를 바탕으로 국가별 기술의 진보 및 혁신의 패턴을 비교한다. 특히, 기술 분야 별 특허등록수를 조사하여, 국가별 집중 기술 분야를 파악하고 군집분석을 이용하여 국가 간 기술진보 유사성을 파악한다. 4장에서는 기술수준의 질적 평가를 할 수 있는 분석 방법론을 소개한다. 단순히 등록특허의 수 뿐 아니라 등록특허의 질을 평가하는 방법을 개발하고 이를 적용하여 각 기술 분야 별 질적 평가를 수행한다. 특히, 특허인용정보를 바탕으로 특허의 중요도를 측정할 수 있는 다양한 네트워크 분석방법론을 소개한다. 5장에서는 각 국가의 사회발전을 파악할 수 있는 다양한 지표를 조사하고 이를 바탕으로 국가별 사회발전 패턴을 파악하며, 정준상관분석을 이용하여 특허자료로부터 파악한 국가별 기술진보 패턴과 국가별 사회발전 패턴과의 상관성을 분석하여 기술진보가 사회발전에 미치는 영향을 정량적으로 파악한다.

2. 특허법 체계의 변화와 영향 분석

근대적인 특허제도는 15세기부터 특정 제품의 생산을 촉진하기 위해서 베니스와 영국 등에서 도입되기 시작했다. 과연 특정기술 또는 제품에 대해서 독점적인 판매 등 특권을 부여하는 것이 기술혁신에 효율적인 법제도인지에 대해서는 많은 논란이 제기되어 왔으나, 독점의 기능성만으로 특허제도가 기술혁신에 방해가 된다고나 연구개발을 위축시킨다고 말하기는 어렵다 (Scherer와 Weisburst, 1995). 나아가, 현실적으로 입증하기 어렵지만, 특허제도가 발명과 혁신을 유인·장려해서 과학·기술 및 산업과 경제의 발전에 기여하기 위한 법제도라고 평가되고 있다 (Posner, 1977).

어떠한 발명과 혁신에 대하여 어떠한 특허권을 부여할 것인가는 특정 국가의 경제적 발전단계와 기술수준에 따라서 좌우될 수 있다. 예컨대 특허 받을 수 있는 발명으로서 화학물질 등의 물질발명이나 유전공학적인 발명이나 소프트웨어발명에 대해서 특허권을 부여할 것인가 등의 문제는 특정 국가의 경제적 발전 단계와 기술수준에 따라서 당해 국가가 법정정책적으로 결정할 문제인 것이다. 세계 각국의 경험들을 종합해 보면, 특정 국가의 1인당 국민소득이 미화 8,000달러에 이를 때까지는 당해 국가의 특허법제도가 아주 낮은 수준의 보호만을 인정하고, 그 이상으로 국민소득이 증가함에 따라서 특허권의 보호수준을 높이게 되는 현상을 볼 수 있다 (Barton, 2002).

우리나라의 경우에도, 화학산업 또는 의약산업의 후진성으로 인하여 물질발명에 대한 특허법적 보호를 인정해 오지 않다가 1987년 7월 1일부터 물질발명에 대한 특허법적 보호를 인정하기 시작하였다. 우리 경제의 중심이 경공업에서 중화학공업 중심으로 그리고 노동집약산업으로부터 기술집약산업으로 이동함에 따라서, 물질발명에 대한 특허법적 보호가 필요하다는 정책적 판단이 이루어지게 되었고, 결과적으로 물질발명에 대한 특허법적 보호가 화학물질의 자체개발을 위한 연구개발의 증가와 국내기업의 물질발명에 관한 특허출원의 증가 그리고 외국화학업체들의 국내 직접투자의 증가라고 하는 긍정적 결과를 가져다주게 되었다.

특허출원, 특허등록, 기술인용 등의 특허정보를 보면 R&D투자, 기술혁신, 제도변화 및 사회변화를 확인해보고 향후 변화방향을 짐작해볼 수 있다. 다만, 특허출원, 특허등록, 기술인용 등 특허활동의 증가 또는 변화가 기술 혁신만에 의해서 영향 받은 것이라고 단정할 수는 없다. 특허활동의 증가는 기술혁신 뿐만 아니라 특허법의 개정, 특허법원의 설립, 특허판례의 변화, 출원 대비 등록 확률과 출원료 등에 의

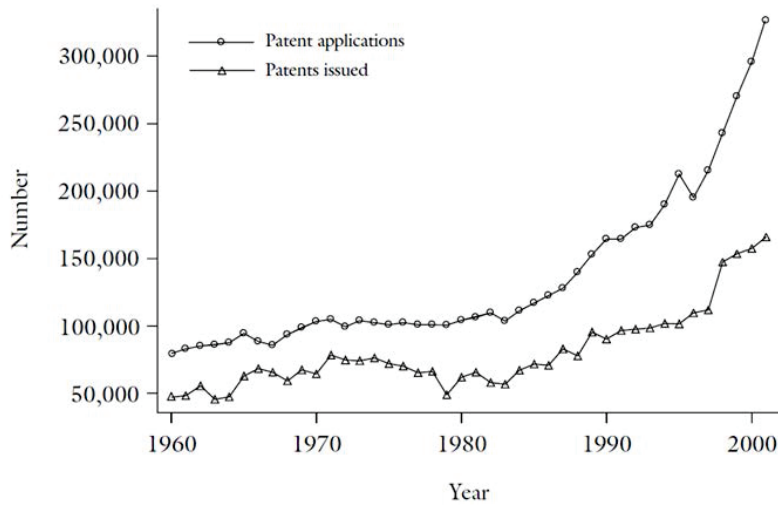


Figure 2.1. The numbers of patents applications and registrations in US, 1960–2001.

해서도 크게 영향받을 수 있기 때문이다 (Landes와 Posner, 2004). 아래에서는 특허법률체계, 특허활동 그리고 기술혁신에 대한 상호 관계 및 영향에 대해서 간단히 살펴본다.

첫째, 기술혁신과 특허활동의 상호관계가 각 산업분야마다 조금씩 다르다. 예컨대, 항공산업, 국방산업, 자동차산업 등의 경우에는 무수히 반복되는 테스트 비중이 높고 특허취득의 비율은 상대적으로 낮다. 정 반대로 기계 산업이나 장치산업의 경우에는 아주 조그만 발명도 특허취득으로 연결되지만, 특정 기계 산업 내에서 특허취득의 건수가 많다고 해서 그렇지 못한 항공 산업에서보다 아주 뛰어난 기술혁신이 이루어졌다고 단언하기는 어렵다 (Pavitt, 1982).

둘째, 특허법원과 연방대법원의 판례가 초래한 특허법의 커다란 변화 또한 특허출원과 등록의 증가에 상당한 기여하였다 특히, 유전공학과 소프트웨어 분야의 특허출원 및 등록건수의 증가는 그러한 분야에서의 특허보호대상의 명확화 내지 확대를 분명히 한 판례로 인한 것이라고 단언할 수 있다. 따라서, 유전공학산업과 소프트웨어산업과 같이 일정분야의 산업은 특허판례에 의해서 기술혁신이 촉진되었다고 하는 점을 부인할 수 없다.

셋째, 기술혁신과 특허활동의 상호관계가 연도별로 상당한 차이를 보이고 있는데, 특히 1982년도부터 미국에서의 특허활동이 크게 증가한 것은 특허법원의 설립에 의해서 영향 받았다는 점을 부인할 수 없다. Figure 2.1의 미국에서의 특허출원 및 등록에 관한 위의 도표를 보면 1982년을 분기점으로 해서 커다란 증가추세를 보여주고 있다. 특히, 특허제도에 대해서 호의적인 입장을 가진 미국연방특허법원(Court of Appeals for the Federal Circuit)이 설립된 1982년도부터 특허출원건수는 5.7%의 아주 급격한 증가를 보여주게 된다 (Posner와 Landes, 2003).

넷째, 특허활동 가운데 특허소송의 증가도 기술혁신과 비례관계를 가진다고 볼 수 있다. Figure 2.2를 보면, 일반 민사소송건수의 변화추세와는 전혀 달리 미국에서의 특허소송건수, 특허변호사, 변리사의 수가 지속적으로 증가해 온 것은 기술혁신이 주된 원인이라고 추측케 한다. 다른 한편, 특허소송의 급증이 특허출원 및 등록건수 증가로 인한 당연한 결과일뿐이라고 하는 견해도 있다. 즉, 특허소송의 증가가 특허법제도의 변화 및 특허법원의 설립으로 인해서 특허출원 등의 특허활동이 급증했기 때문이라고 보는 견해도 있다 (Posner와 Landes, 2003).

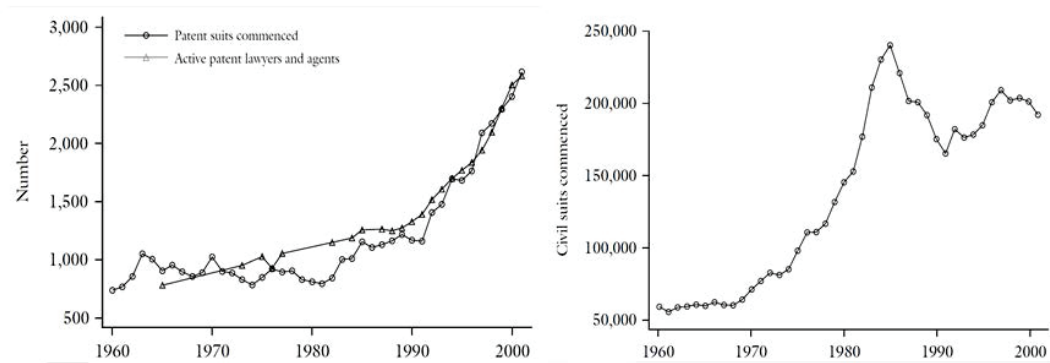


Figure 2.2. Patent cases filed in federal district courts/patent attorneys and agents registered, 1960–2001 (left), Civil cases filed in federal district courts, 1960–2001 (right).

3. 군집분석을 이용한 국가별 기술혁신과 흐름 분석

기술혁신은 국가의 경제적 발전뿐 아니라 사회/문화발전에도 크게 영향을 미친다. 하지만 기술혁신이 국가의 경제/사회/문화의 발전에 어떤 역할을 하였는지를 종합적으로 이해하기 위해서는 기술혁신에 대한 개관적인 지표의 선정이 필수적이다. 기술혁신의 측정 지표로는 크게 R&D투자액, 특허등록수, 그리고 기술혁신에 대한 학술논문수 등을 고려할 수 있다 (Acs 등, 2002). 하지만, R&D 투자액은 정확한 산출이 어려우며, 학술논문은 작은 회사들의 기술혁신을 반영하지 못하는 단점이 있다. 이에 비하여, 특허등록수는, 많은 반대의견도 있지만 (Griliches, 1979; Pakes와 Griliches, 1980; Hall 등, 2001), 기술혁신에 대한 객관적인 지표로써 유용하게 사용될 수 있다는 것이 다양한 논문에서 설명되고 있다 (Griliches, 1990; Crepon과 Duguet, 1997; Acs 등, 2002; Duguet와 Macgravie, 2005).

본 절에서는 미국 등록 특허데이터베이스 자료를 이용하여 국가별로 기술혁신의 패턴을 파악하고, 이를 바탕으로 지역적/문화적으로 기술혁신의 패턴이 어떻게 연결되어 있는지를 파악하고자 한다. 즉, 문화적/지역적으로 상관이 높은 국가들의 기술혁신 패턴도 비슷한지에 대한 실증적 분석을 통하여, 지역/문화적 특징이 기술혁신에 미치는 영향을 파악하고자 한다. Bottazzi와 Peri (2003)는 기술혁신의 지역적 확산에 대하여 연구를 하였는데, 소지역 단위로 분석을 시행하였다. 본 연구에서는 국가단위로 기술혁신의 패턴을 분석하여, 지역적인 특성 이외에 문화적 특성이 기술혁신에 미치는 영향도 살펴보고자 한다.

본 연구는 다음과 같이 진행된다. 미국 특허청 자료를 바탕으로 주요 특허등록국가 15개를 선정하고, International Patent Classification(IPC) level 1에서 주요 기술군을 선정한 후, 국가별/기술군별 특허등록수를 바탕으로 기술혁신의 패턴을 조사하고 유사한 기술혁신패턴을 보이는 국가들의 군집을 생성한 후, 각 군집의 지리적/문화적 유사성을 파악한다.

본 연구에서 사용한 데이터는 미국 특허청에 등록된 특허 리스트와 그 특허들의 기술 분야 정보, 그리고 특허들의 발명자 정보이다. 여기서 기술 분야 정보는 IPC 기준으로 하였으며, 2개 이상의 기술 분야를 가지는 특허에 대해서는 첫 번째로 쓰여 있는 main IPC를 특허의 기술 분야로 정의하였고 (Table 3.1), 특허의 국적은 해당 특허의 발명자의 국적으로 정의하였다. 비교적 특허의 등록 수가 많은 1985년부터 2012년까지 등록된 특허만을 분석 대상으로 삼았고, 이 때 등록된 특허의 수는 4,348,998건이다. 마찬가지로 특허의 발명자 자료 또한 완벽하게 정리되어있지 않기 때문에 발명자의 국적을 알 수 없는 자료

Table 3.1. Technical fields of IPC level 1

IPC level 1	Technical areas
A	Human necessities
B	Performing operations, Transporting
C	Chemistry, Metallurgy
D	Textiles, Paper
E	Fixed constructions
F	Mechanical engineering, Lighting, Heating, Weapons, Blasting
G	Physics
H	Electricity

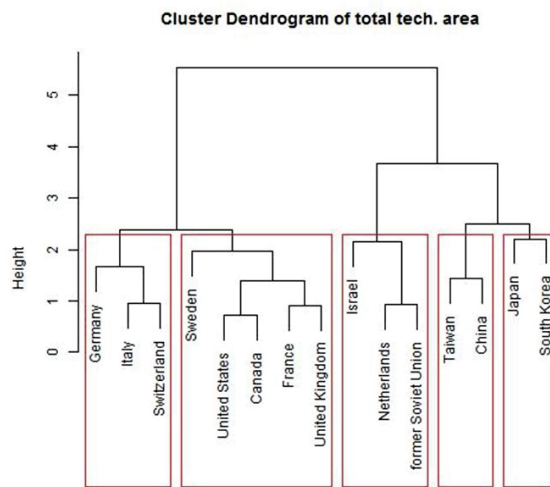


Figure 3.1. Results of hierarchical clustering analysis based on the ratios of registered patents of technical fields of IPS level 1 in each country.

가 다수 존재하여 1985년부터 2012년까지 등록된 특허 중에 발명자들과 발명자들 각각의 국적을 모두 알 수 있는 특허 2,727,957건을 대상으로 분석을 수행하였다.

국가간 기술혁신 패턴 비교를 위하여 선정된 15개의 국가는 1985년부터 2012년 사이에 미국 특허청에 가장 등록을 많이 한 국가 14개(미국, 일본, 독일, 한국, 프랑스, 대만, 영국, 캐나다, 이탈리아, 스위스, 네덜란드, 스웨덴, 이스라엘, 중국)와 등록 특허 개수와는 무관하게 분석해보고 싶은 대상 1개(구소련)로 구성되어 있다. 특히, 국가별 기술군 점유율을 바탕으로 IPC level 1에서 계층적 군집 분석을 실시하여 15개의 국가들을 몇 개의 군집으로 묶어서 군집별 특징을 알아보았고, 국가별 주요 기술혁신 분야의 IPC level 2을 조사하여 자세한 기술혁신 패턴을 국가별로 비교하였다.

국가별로 IPC level 1에서의 등록 특허의 비율을 이용하여 계층적 군집 분석을 실시한 결과는 Figure 3.1에 나타나있다. 계층적 군집 분석에서 각 나라들의 값은 기술군들의 비율을 이용한 8차원 벡터 값을 사용하였으며, 거리는 가장 일반적인 유클리디안 거리를 사용하였고, 이를 토대로 15개의 국가들을 5개의 군집으로 묶어보았다. 그리고 이 결과를 토대로 군집별로 IPC level 1에서의 등록 특허 비율을 나타낸 그림은 Figure 3.2에서 확인할 수 있다(파이 차트에서의 군집 순서는 덴드로그램에서 묶여져 있는 군집 순서와 동일하다).

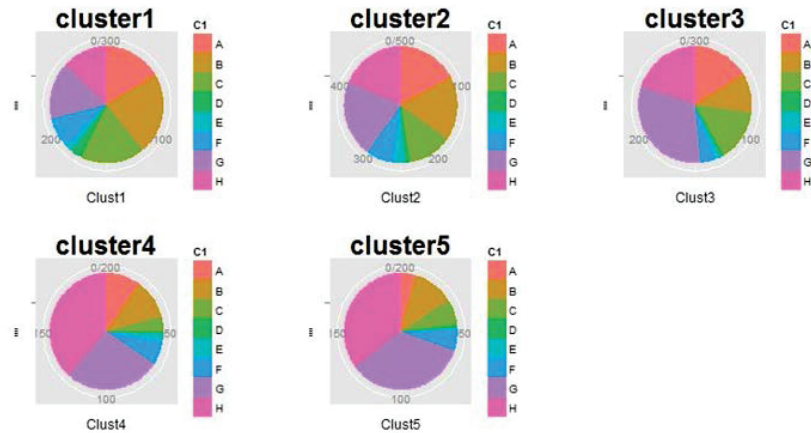


Figure 3.2. The pie charts representing the numbers of registered patents of the 8 technical fields of IPC level 1 in each cluster.

첫 번째와 두 번째 군집은 미국과 유럽 국가들로 이루어져 있는데 모든 기술군에서의 특허 등록 비율이 균형 잡혀 있다는 것을 확인할 수 있다. 그에 반해, 4번째 군집과 5번째 군집은 G 기술과 H 기술의 특허의 비율이 매우 높다는 사실을 볼 수 있는데, 이를 통해 미국 및 유럽 국가들은 거의 모든 기술군을 균형 있게 발전시키지만, 아시아 국가는 특정 몇 개의 기술군을 집중적으로 개발한다는 결론을 추론해볼 수 있다. 우리나라와 중국, 대만 등은 경제력을 키우기 위해서 몇 가지의 기술에 초점을 맞추고 집중적으로 산업을 육성시켜왔는데 위의 분석의 결과는 그 흐름과 일치한다고 볼 수 있겠다. 앞에서 언급했던 주요 기술은 IPC level 1에서 A, G, H 기술군에 속하기 때문에 IPC level 1에서의 주요 기술군은 A, G, H로 정의할 수 있고, 나머지 B, C, D, E, F 기술군은 비주류 기술군이라 정의할 수 있다. 이렇게 정의한 2가지의 기술군에서도 각각 계층적 군집 분석을 시행할 수 있는데 이 결과는 방금 전에 한 분석의 결과와 비슷하기 때문에 생략하였다.

IPC level 1은 너무 광범위한 기술 분야들을 포함하고 있기 때문에 이를 이용한 결과는 대략적인 발전 산업 분야는 알 수 있지만 정확히 특정 국가가 어떤 기술 산업을 집중적으로 발전시키고 있는지에 대해 정확히 알기는 힘들다. 따라서 IPC level 2에서의 기술 점유율을 바탕으로 각 국가별로 가장 특허 등록 비율이 높은 기술 분야 2개씩을 알아보았다. Figure 3.2에 결과가 표시되어 있는데 이를 살펴보면 대부분의 국가가 앞에서 언급했던 주요 기술 분야인 G06, H01, H04, A61을 국가의 주요 기술 분야로 삼고 있다는 것을 확인할 수 있다. 전통적으로 선진국이었던 국가들이 A61 기술(의학/수의학) 분야를 매우 중요하게 생각하고 있음을 알 수 있으며, 개발 도상국들은 대체로 G, H 기술군에 집중하고 있음을 볼 수 있다. 한국과 스웨덴 만이 상위 2개의 기술군 중에 H04(통신 기술)가 존재하는데 이는 한국과 스웨덴에 세계적인 대형 통신사(예를 들어, 한국에는 SK와 LG, KT 등이 있고, 스웨덴에는 에릭센 등이 있다.)가 존재하기 때문인 것으로 생각해볼 수 있다. 또한, 이탈리아와 스위스에 C07(유기 화학) 기술 분야가 상위 2번째에 표시되어 있는데 이는 앞에서 군집분석을 할 때 1번째 군집에서 C의 비율이 다른 군집들과 비교할 때 월등히 크다는 결과와 일치한다는 것도 알 수 있다 (Table 3.2).

4. 네트워크 분석을 이용한 특허 인용 정보에서의 중요도 분석

특허 인용 분석이란 특허 간 인용을 이용하여 만들어진 네트워크를 분석하는 것을 의미한다. 특허가 출

Table 3.2. The two technical fields of IPC level 2 having the most registered patents in each country

Nation	Top1 C2	Area description	Top2 C2	Area description
United States	H01	Basic electric elements	G06	Computing and calculating
Japan	G06	Computing and calculating	A61	Medical or veterinary science
Germany	A61	Medical or veterinary science	H01	Basic electric elements
South Korea	H01	Basic electric elements	H04	Electric communication
France	A61	Medical or veterinary science	H01	Basic electric elements
Taiwan	H01	Basic electric elements	G06	Computing and calculating
United Kingdom	A61	Medical or veterinary science	G06	Computing and calculating
Canada	H04	Electric communication	G06	Computing and calculating
Italy	A61	Medical or veterinary science	C07	Organic chemistry
Switzerland	A61	Medical or veterinary science	C07	Organic chemistry
Netherlands	H01	Basic electric elements	A61	Medical or veterinary science
Sweden	A61	Medical or veterinary science	H04	Electric communication
Israel	G06	Computing and calculating	A61	Medical or veterinary science
China	G06	Computing and calculating	H01	Basic electric elements
The former Soviet Union	G06	Computing and calculating	A61	Medical or veterinary science

원될 때에는 출원자, 또는 심사관이 다른 특허를 인용하는데 이러한 구조는 특허 인용 네트워크를 생성할 수 있게 한다. 일반적인 네트워크 분석 방법을 이용하여 특허 인용 네트워크에 대한 여러 가지 분석이 가능한데, 본 절에서는 중요한 특허 기술군을 찾는 방법을 소개한다. 특허에는 여러 분야, 즉 여러 기술군이 있지만 서로 다른 중요도를 가지고 있다. 특허 출원자가 특허를 유지하기 위해서는 매년 비용이 들어가는데, 이런 특허의 가치를 판단하기 위해서는 중요한 특허 기술군이 무엇인지 아는 것이 중요하다. 이 분석에서는 Google의 PageRank Algorithm (Page 등, 1999)을 이용하여 중요한 기술군을 찾는 방법을 소개한다. 추가적으로, 네트워크 분석에서 주로 쓰이는 여러 vertex centrality와 어떤 차이가 있는지 알아본다. 본 분석에 대한 구체적인 방법 소개는 Lee 등 (2016)의 내용을 참고하였다.

분석에 사용한 자료는 1985년부터 2012년까지의 특허 인용 자료로, 총 특허의 개수는 4,183,884개, 특허 간 인용 횟수는 35,723,262회이다. 즉 특허 인용 네트워크는 4,183,884개의 vertex, 35,723,262개의 edge로 이루어져 있으며, graph density는 $2.04076e-06$ 으로 매우 작다. 그렇기 때문에 개별 특허를 vertex로 하는 네트워크를 분석하는 것보다 더 낮은 level의 기술군을 vertex로 갖는 네트워크를 분석하는 것이 더 의미있을 수 있다. 또한 개별 특허를 vertex로 갖는 네트워크의 경우, 이전에 출원된 특허는 후에 출원된 특허를 인용할 수 없기 때문에, 쌍방향의 edge가 존재할 수 없다. 하지만 높은 level의 vertex를 갖는 네트워크의 경우, 기술군 간의 인용이 쌍방향으로 이루어질 수 있다. 따라서 이 분석에서는 IPC level 3 기술군을 vertex로 갖는 네트워크를 이용하여 분석하였다.

4.1. PageRank 알고리즘 소개

PageRank 알고리즘은 Google의 설립자인 Larry Page와 Sergei Brin이 고안한 것으로, 웹사이트의 상대적 중요도를 계산하는 알고리즘이다. 이 알고리즘은 웹페이지 간에 서로 링크하는 구조를 이용하는 데, 어떤 i 라는 웹페이지의 중요도, 즉 PageRank는 다음과 같은 개념을 기초로 구하게 된다.

1. i 를 링크한 웹페이지들 중, 높은 PageRank를 갖는 웹페이지에 더 많은 가중치를 부여한다.
2. i 를 링크한 웹페이지들 중, 일반적으로 다른 웹페이지에 링크를 많이 한 웹페이지에 더 적은 가중치를 부여한다.

Table 4.1. The 10 most important technical fields of IPC level 3

	C3	PageRank	내용
1	G06F	0.0325	전기에 의한 디지털 데이터처리
2	H01L	0.0263	반도체장치; 다른곳에속하지않는전기적고체장치
3	B32B	0.0160	적층체
4	A61B	0.0159	진단; 수술; 개인 식별
5	A61K	0.0150	의약품, 치과용 또는 화장용 제제
6	B65D	0.0149	물품 또는 재료의 보관 또는 수송용의 용기
7	G01N	0.0132	재료의 화학적 또는 물리적 성질의 검출에 의한 재료의 조사 또는 분석
8	B01D	0.0120	분리
9	A61M	0.0118	흡인 또는 펌프장치
10	H04N	0.0112	화상통신, 예: 텔레비전

이렇게 순환적인 구조를 갖는 알고리즘은 웹사이트의 네트워크가 특정 구조를 가지고 있을 때 문제가 발생한다. 여기서 특정 구조라는 것은 서로 완전히 동떨어진 요소들을 갖는 구조, 루프 구조 등을 의미한다. 이러한 네트워크에서는 각 웹사이트의 PageRank가 유일하게 정의되지 않을 수 있다는 문제점이 있다. 따라서 PageRank 알고리즘에 random surfer 개념을 도입하게 되는데, 이를 이해하기 위해 웹 서핑을 하는 random surfer가 있다고 가정해보자. 이 사람은 현재 웹사이트에서 다른 웹사이트로 가고자 할 때, 현재 웹사이트에서 링크된 웹사이트 중 하나로 이동할 수도 있고, 링크되지 않은 웹사이트 중 임의로 선택된 하나로 이동할 수도 있다. 이 때, 각 웹사이트의 PageRank는 random surfer가 그 웹사이트에 머무르는 시간에 비례한다고 생각할 수 있다. PageRank 알고리즘에 대한 자세한 사항은 Lee 등 (2016)을 참고바란다.

IPC level 3 기술군으로 vectex를 구성한 경우 하나의 vertex안에 여러 특허가 포함되며, 따라서 두개의 vertex사이에는 상호 citation회수 만큼의 복수의 edge가 존재하게 된다. 이 citation 횟수를 edge의 weight로 생각하면, 이에 따라 weighted edge를 가지는 네트워크를 만들 수 있다. 이러한 weighted edge를 가지는 네트워크는 634개의 vertex와, 165,298개의 edge(loop, 즉 기술군의 self citation 포함)로 구성된다. Weighted edge를 가지는 네트워크의 경우, edge의 weight를 고려한 PageRank 알고리즘으로 중요한 기술군을 찾을 수 있다. 이러한 weighted PageRank 알고리즘으로 찾은 주요 IPC level 3 기술군 중 상위 10개 기술군이 다음 Table 4.1과 같다. PageRank 점수가 클수록 중요한 기술군임을 나타낸다.

4.2. 그 외의 중요도 측정 방법을 이용한 분석

위의 PageRank 알고리즘은 중요한 특허 기술군을 찾기 위해 사용한 것으로, 이는 일반적인 네트워크 분석에서 vertex의 중요도를 알아보기 위해 사용하는 vertex centrality와 비교해 볼 수 있다. 주로 많이 사용되는 centrality에는 closeness, betweenness, eigen-vector centrality가 있으며 이에 대한 자세한 설명은 Kolaczyk과 Csárdi (2014)나 Lee 등 (2016)를 참조바란다.

이 세 가지 centrality의 개념을 살펴보면, eigen-vector centrality가 PageRank 알고리즘과 비슷한 개념을 사용하고 있다는 것을 알 수 있다. 주변의 이웃한 vertex의 중요도에 따라 자신의 중요도가 정해지며, 이러한 순환적인 구조 때문에 eigen-vector 문제를 푸는 방식으로 중요도를 구한다는 점이 비슷하다. 실제로 상위 10개 주요 C3 기술군을 찾을 때, PageRank 알고리즘을 이용한 것과 위 세 가지 centrality를 이용한 것을 비교하면 Table 4.2에서 볼 수 있듯이 eigen-vector centrality와 PageRank 알고리즘은 비슷한 결과를 나타낸다.

Table 4.2. Comparison of various centralities

	1	2	3	4	5	6	7	8	9	10
Closeness (in)	B07B	A01C	B23C	F23J	A22C	B60H	B21B	B41F	A23C	C06B
Closeness (out)	A21C	B44C	E02B	G03C	B21K	B60J	F42B	D05B	F03G	F16G
Betweenness	G03F	A23B	B21B	G03C	A01N	A61K	C07D	B30B	G01S	C09B
Eigenvector	G06F	H04L	H04N	G06K	G11C	H04M	G09G	H04J	H01L	H04Q
PageRank	G06F	H01L	B32B	A61B	A61K	B65D	G01N	B01D	A61M	H04N

Directed graph의 경우 closeness centrality에서 거리를 쥔 때, 기준이 되는 vertex로 들어오는 path인지 나가는 path인지에 따라 두 가지 타입의 centrality가 정의될 수 있다. Table 4.2에서 Closeness (in)은 기준이 되는 vertex로 들어오는 path의 경우이며, Closeness (out)은 기준이 되는 vertex에서 나가는 path의 경우를 나타낸다. 이 두 경우와 Betweenness centrality 모두 PageRank 알고리즘과는 매우 다른 결과를 보이고 있다. Eigen-vector centrality가 PageRank 알고리즘과 개념은 물론 분석 결과가 가장 비슷하다는 것을 알 수 있다.

5. 정준상관분석을 이용한 기술혁신과 사회진보 간의 관계에 관한 연구

우리 사회는 기술혁신을 통해서 지속적으로 발전해왔으며 사회진보는 기술진보와 늘 병행해왔다. 기술혁신의 중요성을 깨달아 많은 사회학자가 기술과 사회 간의 관계에 대해서 연구했다. Merton (1938)이 처음으로 Science, Technology and Society(STS)를 발표했으며 사회학의 새로운 분야를 발견했다. Cohen(1990)이 추가적으로 Merton's Thesis가 왜 1990년대 이전에 주목받지 못했는가에 대해서 논의했다. 이 외에도 많은 학자가 기술과 사회에 대해서 정성적으로 연구했다 (Suchman, 2008). 경제활동 등이 우리에게 가장 중요한 사회활동 중 하나이기 때문에 기술과 경제간의 관계를 정량적으로 연구하는 학자도 많았다 (Fagerberg, 1987; Ruttan, 2000).

우리는 이런 질문을 할 수 있다: 정량적으로 기술과 사회 간의 관계는 어떤가? 본 절에서는 이 문제에 대한 답을 줄 수 있는 자료분석 방법론을 소개한다. 특히, 기술혁신과 사회진보를 각각 정량화한 다음에 둘 간의 관계에 대해서 정준상관분석을 통하여 분석하는 방법론에 대해서 설명한다.

5.1. 데이터

Basberg (1987)가 기술혁신 수준은 특허로 측정할 수 있다고 주장했다. Bottazzi와 Peri (2003)가 특허 데이터로 지역 기술혁신 정도를 측정하여 기술발전 및 기술혁신에 있어서 외부성 효과가 있는지를 확인했다. Niebuhr (2010)는 특허 수로 기술혁신을 측정하여 문화적 다양성이 기술혁신에 어떤 영향을 미치는 지에 대해서 연구했다. 반면에, 사회진보의 정량적 측정을 위하여는 ‘진보된 사회’라는 개념에 대한 정의가 필요하다. 진보된 사회와 인간의 욕구에 관한 이론 중에 가장 널리 알려지고 보편적으로 받아들인 이론이 Maslow's Hierarchy of Human Need Theory이다. 즉, 진보된 사회란 사회구성원이 자신의 모든 욕구를 충족할 수 있는 환경을 제공하는 사회라고 정의한다. 특히, 인간의 욕구를 5개의 범주인 생리욕구, 안전욕구, 애정소속욕구, 존경욕구 그리고 자아실현욕구로 나누었으며, 각각의 범주는 여러 개의 소비주로 나누어 진다.

본 연구에서 특허 데이터로 기술혁신을 측정하고자 한다. 매년 미국특허청에 등록된 특허수로 한 국가의 기술혁신 정도를 측정한다. IPC으로 기술을 분류하여 각 분야의 기술혁신 정도를 알 수 있다. 본 연구에서 IPC level 1 레벨의 기술분류를 사용한다. 인간의 욕구에 대한 Maslow's Theory에 의한 5범주

Table 5.1. Indicators used in Maslow's theory

Maslow's Hierarchy of Need	Measurements
1. Physiological	Life expectancy at birth(LE) Income shared held by lowest 10%(Low)
2. Safety	Homicide rate(HR) Air pollution(AP)
3. Belongingness and Love	Importance of family in life(IOFAM) Importance of friends in life(IOFRI)
4. Esteem	Trust(MPT) Respect of human rights(RHR)
5. Self-Actualization	School enrollment of tertiary(SET) Importance of leisure(IOL) Freedom of life control(FOC)

에 대해서 2개나 2개 이상의 사회지표로 측정할 것이며 사용된 변수는 Table 5.1에 정리되어 있다. 생리욕구에서 출생시기대수명(LE)과 하위 10% 인구가 차지하는 부의 비율(Low)은 World Bank에서 제공된 자료다. 안전욕구에 대한 지표 중 사회안전수준을 나타내는 살인율(HR)은 UN Office on Drugs and Crime에서 자료를 획득할 수 있으며 자연안전수준을 나타내는 대기오염(AP)은 Carbon Dioxide Information Analysis Center에서 제공한 나라별 년 CO2 배출량 데이터를 사용했다. 나머지 3가지 욕구의 경우 개인의 감정에 관한 내용이기 때문에 상황을 더 정확히 반영하기 위해서 World Value Survey에서 제공하는 설문데이터를 사용하기로 했다. 애정소속욕구에서는 가족의 중요도(IOFAM)와 친구의 중요도(IOFRI)를 사용하며 존경욕구에서는 신뢰수준(MPT)과 인권존중정도(RHR)를 사용하며 자아실현욕구에서는 여가의 중요도(IOL)와 삶의 컨트롤 정도(FOC)를 사용 한다. 자아실현욕구에서는 추가적으로 대학진학율(SET)를 사용하며 이는 UN Educational, Scientific and Cultural Organization(UNESCO)에서 제공한 데이터를 사용했다.

5.2. 정준상관분석

본 연구는 기술혁신과 사회진보를 각각 지표화한 다음 둘 간의 상관관계를 정량적으로 분석하고자 한다. 특히 데이터로 기술혁신을 측정하고 11개 사회지표로 사회진보를 측정할 것이다. 따라서 본 연구는 두 변수군의 상관관계를 분석하는 것이다. 사회과학의 특성때문에 많은 연구가 다변량 분석을 해야 하며, 정준상관분석(canonical correlation analysis; CCA)으로 두 변수군의 상관관계를 분석할 수 있다 (Onwuegbuzie와 Daniel, 2003). 정준상관분석은 Hotelling (1936)이 처음으로 제시한 것인데, 정준상관분석의 목적은 변수군의 선형조합으로 두 변수군의 상관관계 구조를 분석하는 것이다. 두 변수군 \mathbf{X} 와 \mathbf{Y} 가 있으면 공분산 행렬을 다음과 같이 정의할 수 있다.

$$\text{Cov} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

\mathbf{X} 와 \mathbf{Y} 의 선형조합을 $\mathbf{a}'\mathbf{X}$ 와 $\mathbf{b}'\mathbf{Y}$ 로 정의한다. 따라서 $\mathbf{a}'\mathbf{X}$ 와 $\mathbf{b}'\mathbf{Y}$ 의 공분산은 $\text{Cov}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y}) = \mathbf{a}'\Sigma_{12}\mathbf{b}$ 이며 상관계수는:

$$\text{Corr}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y}) = \frac{\mathbf{a}'\Sigma_{12}\mathbf{b}}{(\mathbf{a}'\Sigma_{11}\mathbf{a} \cdot \mathbf{b}'\Sigma_{22}\mathbf{b})^{-\frac{1}{2}}}. \quad (5.1)$$

Table 5.2. Statistical significance of the relation between physiological need and technical innovation

rho	Sq rho	Chi-square	df
0.4447	0.1977	36.5075	16

정준상관분석은 식 (5.1)을 최대화하는 \mathbf{a} 와 \mathbf{b} 를 찾는 것이다. 두 선형조합은 $U_1 = \mathbf{a}'_1 \mathbf{X} = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$ 과 $V_1 = \mathbf{b}'_1 \mathbf{Y} = b_{11}y_1 + b_{12}y_2 + \dots + b_{1q}y_q$ 로 정의한다. 그리고 $\text{Var}(U_1) = \text{Var}(V_1) = 1$, $\rho(U_1, V_1) = \max_{\mathbf{a}_1, \mathbf{b}_1} \rho(\mathbf{a}'_1 \mathbf{X}, \mathbf{b}'_1 \mathbf{Y})$ 일 때, (U_1, V_1) 은 첫 번째 정준상관변수쌍이라고 하며 $\rho_1^* = \max_{\mathbf{a}, \mathbf{b}} \rho(\mathbf{a}' \mathbf{X}, \mathbf{b}' \mathbf{Y})$ 는 첫 번째 정준상관계수라고 부른다. 그리고 $U_2 = \mathbf{a}'_2 \mathbf{X}$, $V_2 = \mathbf{b}'_2 \mathbf{Y}$ 일 때, $\text{Cov}(\mathbf{a}'_2 \mathbf{X}, U_1) = \text{Cov}(\mathbf{b}'_2 \mathbf{Y}, V_1) = 0$, $\text{Var}(U_2) = \text{Var}(V_2) = 1$ 인 $\mathbf{a}_2, \mathbf{b}_2$ 에 대하여 $\rho(\mathbf{a}'_2 \mathbf{X}, \mathbf{b}'_2 \mathbf{Y})$ 를 최대로 하는 (U_2, V_2) 를 두 번째 정준상관변수쌍이라고 하며 이 때의 $\rho(\mathbf{a}'_2 \mathbf{X}, \mathbf{b}'_2 \mathbf{Y})$ 값을 두 번째 정준상관계수라고 부른다. 첫 번째 정준상관변수쌍이 두 변수군의 선형조합에서 가장 큰 정준상관관계 제곱을 나타내는 것이며, 두 번째 정준상관변수쌍은 두 번째로 큰 정준상관관계 제곱을 나타내며 첫 번째 정준상관변수쌍과 독립이다. 비슷한 논리로 두 변수군의 k 번째 선형조합 (U_k, V_k) 는 k 번째 큰 정준상관관계 제곱이며 $(U_1, V_1), \dots, (U_{k-1}, V_{k-1})$ 과 모두 독립이다.

첫 번째 정준상관변수쌍은 다음과 같은 선형조합을 통해서 구할 수 있다:

$$U_1 = \mathbf{e}'_1 \Sigma_{11}^{-\frac{1}{2}} \mathbf{X} \quad \text{and} \quad V_1 = \mathbf{f}'_1 \Sigma_{22}^{-\frac{1}{2}} \mathbf{Y} \tag{5.2}$$

$\rho_1^{*2} \geq \rho_2^{*2} \geq \dots \geq \rho_p^{*2}$ 은 $\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2}$ 의 고유값(eigenvalues)이며, $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$ 는 각각의 고유값에 대응하는 고유벡터(eigenvectors)이다. 또한, $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_q$ 는 $\Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1/2} \mathbf{e}_1$ 에 비례한다. 정준상관분석의 특성에 따라서 두 변수군 중에 변수의 개수가 적은 변수군의 변수 개수만큼 정준상관을 구할 수 있다. 그러나 모든 정준상관을 해석하는 것이 의미가 없다. 공분산을 충분히 설명하는 정준상관계수를 해석하면 된다. 하지만 ‘충분한’ 공분산에 대한 정의는 없기 때문에 몇 쌍의 정준상관계수를 해석하는 것은 연구자의 몫이다. 본 연구에서는 편의상 첫 번째 정준상관만 분석한다.

5.3. 분석결과

7개 국가(Argentina, India, Mexico, Poland, Spain, Sweden, United States)로부터 데이터를 구하고, 이를 합쳐서 분석에 사용하였다. 7개 국가 중에 선진국과 개발도상국이 있으며 아시아, 유럽, 북미와 남미 모두 있으므로, 본 분석의 결과는 다양한 국가/사회에 적용될 수 있을 것이다.

생리욕구관련 변수군과 기술혁신변수 간의 정준상관분석의 통계적 유의성은 Table 5.2에 정리하였다. 표에서 ‘rho’는 첫 번째 고유값 ρ_1^* 의 추정값이고 ‘Sq rho’는 고유값의 제곱, 그리고 ‘Chi-square’와 ‘df’는 첫 번째 정준상관의 통계적 유의성 검정을 위한 통계값이다. 여기서 ‘Sq rho’는 회귀분석의 R^2 과 비슷한 역할을 하며 전체 상관관계 중 첫 번째 정준상관변수가 설명하는 양이라고 해석할 수 있다. 유의성 검정에 의하면 첫 번째 정준상관관계는 유의하고($\chi^2_1(16) = 36.51, p < .005$) 전체 상관관계중 19.78%가 첫 번째 정준상관변수로 설명이 된다. 하지만, 일반적으로 ‘Sq rho’로 측정되는 정준상관정도가 0.30 이상일 때만 의미가 있다고 해석이 되며 (Capraro와 Capraro, 2001), 따라서 생리욕구는 기술혁신과 연관성이 없다고 결론을 도출할 수 있다.

Table 5.3은 안전욕구관련 변수군과 기술혁신변수 간의 정준상관분석의 결과이다. 전체 모델은 유의하며($\chi^2_1(16) = 320.02, p < .000$) 첫 번째 정준상관변수가 전체 관계 중 94.92%를 설명하고 있다. 따라서, 안전욕구와 기술혁신간에는 유의한 관계가 있다고 결론을 내릴 수 있다. 즉, 인간의 안전욕구가 관련 기술의 진보와 혁신을 추동하는 원인이다.

Table 5.3. Statistical significance of the relation between safety need and technical innovation

rho	Sq rho	Chi-square	df
0.9742	0.9492	320.0241	16

Table 5.4. CCA Result for the relation between safety need and technical innovation

	Technology Innovation			Social Progress	
	Std. Function Coefficients			Std. Function Coefficients	
	Function 1	Function 2		Function 1	Function 2
A	1.43546550	13.5338134	HR	0.004563188	-1.0070655
B	0.13175129	-5.4528301	AP	-1.000529714	0.1146389
C	-1.63624583	-5.8613633			
D	0.06755221	9.1734996			
E	0.36135516	-20.9217448			
F	-1.03151456	0.2826712			
G	-2.17003383	-4.2507215			
H	1.77551998	14.0271442			
	Structure Coefficients			Structure Coefficients	
	Function 1			Function 2	
	Function 1	Function 2		Function 1	Function 2
A	-0.9776384	0.056159544	HR	-0.1138334	-0.993499847
B	-0.9840759	0.032422367	AP	-0.9999897	-0.004531126
C	-0.9801798	0.009648073			
D	-0.9455690	0.041136306			
E	-0.9864888	0.040641585			
F	-0.9860082	0.033864404			
G	-0.9456770	0.104285030			
H	-0.9685504	0.098376036			
	Sq. Structure Coefficients			Sq. Structure Coefficients	
	Function 1			Function 2	
	Function 1	Function 2		Function 1	Function 2
A	9.56E-01	3.15E-03	HR	0.01295805	0.987041946
B	9.68E-01	1.05E-03	AP	0.99997947	2.05311E-05
C	9.61E-01	9.31E-05			
D	8.94E-01	1.69E-03			
E	9.73E-01	1.65E-03			
F	9.72E-01	1.15E-03			
G	8.94E-01	1.09E-02			
H	9.38E-01	9.68E-03			

첫 번째 정준상관 변수들의 계수를 조사해 보면 각 변수들 사이의 관계를 파악할 수 있다. 정준상관변수의 계수를 분석할 때 주로 표준정준계수(standardized function coefficient), 정준부하량(structure coefficient)과 정준부하량제곱(squared structure coefficient)을 사용한다. 표준정준계수는 두 변수군의 상관관계를 최대화 시키는 계수들의 표준화된 값이다 (Frederick, 1999). 정준부하량은 개별변수와 정준변수간의 단순 상관관계다. 정준부하량제곱은 정준부하량의 제곱이며 개별변수가 정준변수의 분산을 어느정도 설명하는지를 나타내는 계수다. 정준부하량제곱의 합이 1을 넘으면 같은 변수군 간의 변수들끼리 다중공선성이 존재하다는 증거다. 그리고 표준정준계수와 정준부하량의 절대값이 모두 0.30을 넘은 변수만 의미있는 것이며 부호의 방향성은 결과해석할 때 추가적으로 고려해야 한다 (Tabachnick과 Fidell, 2007). Table 5.4에서 나타난 바와 같이 6개 기술분야의(A, B, C, E, F, H) 정준부하량 절대

값이 0.95보다 크며 2개 기술분야의(D, G) 정준부하량 절대값 또한 0.90보다 크다. 이는 8개 기술분야 간의 다중공선성이 매우 강하다는 것이다. HR와 AP의 정준부하량이 모두 음수인데 같은 부호의 정준부하량은 양의 상관관계로 해석할 수 있다. AP와 HR의 표준정준계수는 각각 -1.0005 와 0.0046 이다. HR의 표준정준계수의 절대값이 0.30보다 작기 때문에 HR가 기술혁신과 연관성이 없다고 볼 수 있다. G의 표준정준계수의 절대값이 가장 크며(-2.17) 잇따른 기술분야가 C(-1.64)와 F(-1.03)이다. 그러므로 AP는 G, C, F와 양의 상관관계를 가지며 H(1.78)와 A(1.44)는 AP와 음의 상관관계를 가진다. D(0.07)와 B(0.13)는 2단계인 안전육구와 연관성이 없다고 할 수 있다.

나머지 부분의 결과는 지면의 부족으로 간단하게 정리하면 다음과 같다. 대기오염(AP) 이 물리학(G), 화학 및 금속(C), 기계공학, 조명, 가열 및 무기(F) 기술과 양의 상관관계가 있으며 전기(H), 생활필수품(A) 과는 음의 상관관계가 나타났다. 애정소속육구에서 가족의 중요도(IOFAM)가 기술혁신과 연관성이 없지만 친구의 중요도(IOFRI)는 직물 및 제지(D), 생활필수품(A), 기계공학, 조명, 가열 및 무기(F) 기술과 양의 상관관계를 가지며 고정구조(E), 화학 및 금속(C) 기술과 음의 상관관계를 가진다. 자아실현육구에서 대학진학율(SET) 만 기술혁신과 연관성이 있으며 작업 및 운송(B), 물리학(G), 기계공학, 조명, 가열 및 무기(F) 기술과 양의 상관관계를 가지며 화학 및 금속(C), 전기(H), 고정구조(E) 기술과 음의 상관관계를 가진다.

6. 결론

본 논문에서는 특허빅데이터를 분석하여 기술혁신과 사회변화의 관계를 조망할 수 있는 다양한 분석 방법을 소개하였다. 특히, 특허빅데이터 분석에 군집분석, 네트워크 분석 그리고 정준상관분석 등이 어떻게 사용되어질 수 있는지를 간단하게 설명하였다. 지면의 제약상 주요 아이디어만 소개했는데, 자세한 방법은 추후에 다른 논문들에서 소개하기로 한다.

본 논문에서 소개한 방법 외에 특허빅데이터를 분석하여 유용한 정보를 얻는 방법은 무수히 많을 것이다. 유망기술 발굴, 발명자 네트워크 분석, 특허 키워드 분석, 분쟁특허 예측 등은 많은 분야에서 관심을 가지고 있는 주제들이며, 특허빅데이터를 분석하여 이러한 주제들에 대한 답을 제공할 수 있을 것이다.

References

- Acs, Z., Anselin, L. and Varga, A. (2002). Patents and innovation counts as measures of regional production of new knowledge, *Research Policy*, **31**, 1069–1085.
- Adelman, D. E. and Deanglis, K. L. (2007). Patent metrics: the mismeasure of innovation in the biotech patent debate, *Texas Law Review*, **85**, 1677.
- Barton, J. H. (2002). *Integrating intellectual property rights and development policy: Report of the commission on intellectual property rights*, Commission on Intellectual Property Rights.
- Basberg, B. L. (1987). Patents and the measurement of technological change: a survey of the literature, *Research Policy*, **16**, 131–141.
- Bottazzi, L. and Peri, G. (2003). Innovation and spillovers in regions: evidence from European patent data, *European Economic Review*, **47**, 687–710.
- Capraro, R. M. and Capraro, M. M. (2001). Commonality analysis: Understanding variance contributions to overall canonical correlation effects of attitude toward mathematics on geometry achievement, *Multiple Linear Regression Viewpoints*, **27**, 16–23.
- Cohen, I. B. (1990). *Puritanism and the Rise of Modern Science: the Merton Thesis*, Rutgers University Press, New Brunswick, NJ.
- Crepon, B. and Duguet, E. (1997). Estimating the innovation function from patent numbers: GMM on count panel data, *Journal of Applied Econometrics*, **12**, 243–263.

- Duguet, E. and Macgravie, M. (2005). How well do patent citation measure flows of technology? Evidence from French innovation surveys, *Economics of Innovation and New Technology*, **14**, 375–393.
- Fagerberg, J. (1987). A technology gap approach to why growth rates differ, *Research Policy*, **16**, 87–99.
- Frederick, B. N. (1999). Partitioning variance in the multivariate case: a step-by-step guide to canonical commonality analysis, *Advances in Social Science Methodology*, **5**, 305–318.
- Griliches, Z. (1990). *Patent statistics as economic indicators: a survey* (working paper No.3301), National Bureau of Economic Research.
- Griliches, Z. (1979). Issues in assessing the contribution of R&D to productivity growth, *Bell Journal of Economics*, **10**, 92–116.
- Hall, G., Jaffe, A., and Trajtenberg, M. (2001). *The NBER patent citations data file: lessons, insights and methodological tools* (WP 8498), National Bureau of Economic Research.
- Hotelling, H. (1936). Relations between two sets of variants, *Biometrika*, **28**, 321–377.
- Kerr, S. P. and Kerr, W. R. (2014). *Global collaborative patents* (working paper), National Bureau of Economic Research.
- Kolaczyk, E. and Csárdi, G. (2014). *Statistical Analysis of Network Data with R*. Springer, New York.
- Landes, W. M. and Posner, R. A. (2004). An empirical analysis of the patent court, *The University of Chicago Law Review*, 111–128.
- Lee, M., Kim, Y., and Jang, W. (2016). Patent citation network analysis, *The Korean Journal of Applied Statistics*, **29**, 613–625.
- Merton, R. K. (1938). Science, technology and society in seventeenth century England, *Osiris*, 360–632.
- Niebuhr, A. (2010). Migration and innovation: does cultural diversity matter for regional R&D activity?, *Papers in Regional Science*, **89**, 563–585.
- Onwuegbuzie, A. J. and Daniel, L. G. (2003). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education*, 6.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.
- Pakes, A. and Griliches, Z. (1980). Patents and R&D at the firm level: a first report, *Economic Letters*, **5**, 377–381.
- Pavitt, K. (1982). R&D, patenting and innovative activities: a statistical exploration, *Research Policy*, **11**, 33–51.
- Posner, R. A. (1977). *Economic Analysis of Law*, Little, Brown and Company, Boston and Toronto.
- Posner, R. A. and Landes, W. M. (2003). *The Economic Structure of Intellectual Property Law*, Harvard university press.
- Ruttan, V. W. (2000). Technology, growth, and development: an induced innovation perspective, *OUP Catalogue*.
- Scherer, F. M. and Weisburst, S. (1995). Economic effects of strengthening pharmaceutical patent protection in Italy, *IIC-International Review of Industrial Property and Copyright Law*, **26**, 1009–1024.
- Suchman, L. (2008). Feminist STS and the sciences of the artificial, In *New Handbook of Science and Technology Studies*, MIT Press.
- Tabachnick, B. G. and Fidell, L. S. (2007). Multivariate analysis of variance and covariance, *Using Multivariate Statistics*, **3**, 402–407.

기술의 진보와 혁신, 그리고 사회변화: 특허빅데이터를 이용한 정량적 분석

김용대^{a,1} · 정상조^b · 장원철^a · 이종수^c

^a서울대학교 통계학과, ^b서울대학교 법과대학, ^c서울대학교 기술경영경제정책과정

(2016년 8월 30일 접수, 2016년 10월 10일 수정, 2016년 10월 10일 채택)

요약

본 논문에서는 특허빅데이터를 분석하여 기술적 혁신과 사회변화의 관계를 규명하는 다양한 방법에 대하여 소개를 한다. 특히, 미국특허청에 1985년부터 2015년까지 등록된 4백만개 이상의 특허자료를 분석하였다. 먼저, 특허법의 변천사를 살펴보고 특허법의 발전이 특허활동에 미치는 영향에 대해서 살펴보았다. 두 번째로는, 국가별 기술군별 등록특허수를 바탕으로 군집분석을 이용하여 기술혁신 패턴이 비슷한 국가들로 군집을 만들고 각 군집의 기술혁신 특징들을 살펴보았다. 세 번째로는 특허간의 인용정보를 바탕으로 특허간의 네트워크를 구축하고 page-rank 알고리즘을 이용하여 주요특허를 탐지하는 방법을 설명하였다. 마지막으로, 정준상관분석을 이용하여 기술혁신과 사회변화와의 관계를 규명하였다.

주요용어: 특허, 기술혁신, 군집분석, 네트워크 데이터, 정준상관분석

이 논문은 2014년도 SNU Brain Fusion Program 지원사업의 지원을 받아 수행된 연구임.

¹교신저자: (00826), 서울특별시 관악구 관악로 1, 서울대학교 자연과학대학 통계학과.

E-mail: ydkim903@snu.ac.kr