The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.4.613
29(4), 613–625

Patent citation network analysis

Minjung Lee[a] · Yongdai Kim[b] · Woncheol Jang[b,1]

[a]NAVER; [b]Department of Statistics, Seoul National University

(Received February 11, 2016; Revised April 14, 2016; Accepted April 25, 2016)

Abstract

The development of technology has changed the world drastically. Patent data analysis helps to understand modern technology trends and predict prospective future technology. In this paper, we analyze the patent citation network using the USPTO data between 1985 and 2012 to identify technology trends. We use network centrality measures that include a PageRank algorithm to find core technologies and identify groups of technology with similar properties with statistical network models.

Keywords: patent citation, prospective technology, PageRank algorithm, stochastic block model, latent space model

1. 서론

과학 기술의 흐름과 동향을 나타내는 대표적인 자료 중 하나로 특허 자료를 들 수 있다. 특허 자료 분석 방법은 다양하지만 그 중에서도 특허 인용 네트워크 분석은 특허 간의 인용 관계를 이용하여 만들어진 네트워크를 분석하는 것을 의미한다. 특허를 vertex로, 특허 간 인용을 edge로 간주할 경우 특허 인용 자료를 네트워크 자료로 고려할 수 있다. 특허가 출원될 때에는 출원인이나 심사관이 다른 특허를 인용하는데 출원인은 출원된 발명의 독창성, 진보성을 강조하기 위해 최대한 인용을 하지 않고, 심사관은 관련 특허의 공적을 정확하게 밝히기 위해 특허를 인용한다는 점에서 차이가 있다. 이렇게 출원인 또는 심사관이 관련 특허를 인용하는 구조를 이용하여 특허 인용 네트워크를 생성할 수 있다. 이러한 특허 인용 네트워크를 분석함으로써 특허 간의 상호관계를 파악하여 특허의 가치를 판단하거나 특허 간 연관성을 확인하는 것이 가능하다.

특허 인용 분석에서는 주로 특허의 가치를 평가하기 위한 중심도 지표나 관심있는 기술군의 동향 파악에 관한 분석이 이루어지고 있다. Oh 등 (2012)은 특허 인용의 종류를 구별하고 다른 가중치를 부여함으로써 특허의 순위를 정하는 방법을 소개하였으며, Li 등 (2007)은 나노기술과 관련된 특허를 이용하여 나노기술 연구개발의 현황을 분석하였다. Yoo 등 (2007)은 전기통신 관련 기술군 네트워크를 이용

This work was supported by SNU Brain Fusion Program of the Seoul National University in 2014, the Basic Science Research Program through the National Research Foundation (NRF) of Korea grant by the Ministry of Education (No. 2013R1A1A2010065) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A4A1007895).
[1]Corresponding author: Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-Gu, Seoul 08826, Korea. E-mail: wcjang@snu.ac.kr

Table 2.1. International patent classification (IPC)

Section	Contents
A	Human Necessities
B	Performing Operations, Transporting
C	Chemistry, Metallurgy
D	Textiles, Paper
E	Fixed Constructions
F	Mechanical En gineering, Lighting, Heating, Weapons, Blasting
G	Physics
H	Electricity

해 연관 기술군의 중심도 지표와 기술 지식의 흐름을 분석하였다. 이처럼 특허 인용 네트워크에 대한 다양한 분석이 가능한데, 이 논문에서는 크게 두 가지 분석을 하고자 한다.

첫째, 중요한 특허 기술군을 찾고자 한다. 특허에는 여러 분야, 즉 여러 기술군이 있으며 서로 다른 중요도를 지니고 있다. 특허 출원자가 특허를 유지하기 위해 일정 기간마다 유지비용을 지불해야 하기 때문에 중요한 특허 기술군이 무엇인지 파악함으로써 출원자가 보유한 특허의 가치를 판단하는 것이 필요하다. 또한 시간이 흐르면서 새로운 특허가 출원되거나 기존의 특허가 사라지면서 특허 인용 네트워크는 변화한다. 이 분석에서는 Google의 PageRank 알고리즘 (Brin과 Page, 1998)을 이용해 특허 기술군의 중요도를 계산한다. 추가적으로 네트워크 분석에서 주로 쓰이는 여러 중심도 지표(vertex centrality)와 어떤 차이가 있는지 알아본다. 마지막으로 시대별로 주요 특허 기술군이 어떻게 달라지는지 살펴본다.

둘째, 특허 인용 네트워크에 여러 가지 네트워크 모형을 적합한다. 특허 기술군을 여러 그룹으로 분류할 수 있으며 그에 따라 특허 기술군 간 인용이 발생할 확률이 정해진다는 가정의 stochastic block model과, 관측되지 않는 잠재변수를 고려한 latent space model을 적합해보고 특허 인용 네트워크가 두 모형에 적절한지 분석하고자 한다 (Kolaczyk와 Csárdi, 2014). 네트워크 모형 적합을 통해 기술군의 군집 분석 또한 가능하며 stochastic block model을 통해 얻은 기술군 그룹의 특징과 그룹에 포함된 기술군들의 연관성을 해석해보고자 한다. 또한 분류된 그룹이 latent space model 적합을 통해 얻어진 latent space에서 어떻게 위치하고 있는지 살펴본다.

2. 특허 인용 자료

2.1. 국제 특허기술 분류

특허 인용 자료를 살펴보기 전에 우선 국제 특허 분류 체계에 대한 이해가 필요하다. 국제 특허 분류(international patent classification; IPC)는 특허 문헌을 계층적으로 분류하는 국제적인 기술 분류 체계다. 이는 기술 분야에 따라 특허를 분류하는데, section, class, subclass, group 순서의 계층적 구조로 구성된다. 각 하위분류는 숫자나 알파벳으로 이루어진 기호로 나타낸다. 가장 먼저 8개의 section으로 나뉘며, 그 내용은 Table 2.1과 같다.

Section으로 분류된 후, 하위 분류 체계인 class, subclass, group 순서로 분류된다. Class는 두 자리의 숫자, subclass는 알파벳, group은 숫자로 표시된다. 예를 들어, 터치 스크린 기술은 국제 특허 분류에 따라 'G06F 3/041'라는 기술군으로 분류되는데, 여기서 순서대로 'G'는 '물리학(physics)'을 나타내는 section, '06'은 '산술 논리 연산, 계산, 계수(computing, calculating, counting)'를 나타내는 class, 'F'는 '전기에 의한 디지털 데이터 처리(electric digital data processing)'를 나타내는 subclass, '3/041'은 '변환 수단에 의해 특징지워진 디지타이저(digitisers)'를 나타내는 group을 의미한다.

본 논문에서는 계층적 분류 체계인 section, class, subclass, group의 용어 대신 C1, C2, C3, C4와 같이 간단한 용어를 사용한다.

2.2. 자료 설명

분석에 사용한 자료는 (주) 광개토연구소(https://www.patentpia.com/)에서 제공한 1985년부터 2012년 사이에 미국 특허청(USPTO)에 등록된 특허들 간의 인용 자료로 총 특허의 개수는 3,834,379개, 특허 간 인용 횟수는 32,204,447회이다. 즉 특허 인용 네트워크는 3,834,379개의 vertex, 32,204,447개의 edge로 이루어져 있으며 2.04076e−06이라는 매우 작은 graph density를 갖는다. Graph density는 네트워크에서 연결 가능한 모든 edge 중에서 실제로 연결된 edge의 비율을 나타내는 수치로 이 값이 매우 작다는 것은 개별 특허로 이루어진 네트워크 내에서 특허 간의 인용 관계가 매우 적게 발생한다는 것을 뜻한다. 그렇기 때문에 개별 특허를 vertex로 갖는 네트워크를 분석하는 것보다 더 높은 레벨의 vertex를 갖는 네트워크를 분석하는 것이 더 의미있을 수 있다. 다시 말해서, C3 또는 C2 기술군을 vertex로 갖는 네트워크를 분석하는 편이 더 적합하다. 개별 특허를 vertex로 갖는 네트워크의 경우, 과거에 출원된 특허는 후에 출원된 특허를 인용할 수 없기 때문에 쌍방향의 edge가 존재할 수 없다. 하지만 기술군 간의 인용은 쌍방향으로 이루어질 수 있기 때문에 높은 레벨의 vertex를 갖는 네트워크의 경우 이러한 문제가 생기지 않는다. 따라서 이 분석에서는 C3 또는 C2 레벨의 기술군을 vertex로 갖는 네트워크를 이용하여 분석하였다. 중심도 지표 분석에서는 C3 기술군을 vertex로 갖는 네트워크를 사용했는데, 이는 633개의 C3 기술군(vertex)과 155,813회의 기술군 간 인용(edge)으로 이루어져 있다. 네트워크 모형 분석에서는 군집 분석이 용이하도록 더 작은 네트워크를 사용하였다. 여기에서는 C2 기술군을 vertex로 갖는 네트워크를 사용했는데, 이는 122개의 C2 기술군(vertex)과 13,052회의 기술군 간 인용(edge)으로 이루어져 있다. 또한 출원인의 인용과 심사관의 인용에 차별을 두지 않고 모든 종류의 인용을 사용해 분석하였다.

3. 분석 방법론

3.1. 중심도 지표

3.1.1. PageRank 알고리즘 PageRank 알고리즘은 Google의 설립자인 Larry Page와 Sergei Brin이 고안한 것으로, 웹페이지의 상대적 중요도를 계산하는 알고리즘이다. 이 알고리즘은 웹페이지 간의 링크를 이용하여 웹페이지의 상대적인 중요도를 계산하는데, i라는 웹페이지의 중요도는 i를 링크한 웹페이지의 중요도의 가중합으로 계산하며 다음과 같은 원리를 바탕으로 구하게 된다 (Page 등, 1999).

- i를 링크한 웹페이지들 중, 높은 PageRank를 갖는 웹페이지에 더 많은 가중치를 부여한다.
- i를 링크한 웹페이지들 중, 일반적으로 다른 웹페이지에 링크를 많이 건 웹페이지에 더 적은 가중치를 부여한다.

이렇게 순환적인 구조를 갖는 알고리즘은 웹페이지의 네트워크가 특정 구조를 가지고 있을 때 문제가 발생한다. 여기서 특정 구조라는 것은 서로 완전히 동떨어진 요소들을 갖는 구조, 루프 구조 등을 의미한다. 이러한 구조를 갖는 네트워크에서는 각 웹페이지의 PageRank가 유일하게 정의되지 않을 수 있다는 문제점이 있다. 이러한 문제를 해결하기 위해 PageRank 알고리즘에 random surfer 개념을 도입한다. 이 random surfer는 웹서핑을 하는 사람을 뜻하며, 현재 웹페이지에서 다른 웹페이지로 가고자 할 때 현재 웹페이지에서 링크된 웹페이지 중 하나로 이동할 수도 있고, 링크되지 않은 웹페이지 중 임

의로 선택된 하나로 이동할 수도 있다. 이렇게 링크되지 않은 임의의 다른 웹페이지로 이동할 수 있는 random surfer 개념을 PageRank 알고리즘에 적용하면 앞서 말한 것과 같은 특정 구조를 가지는 네트워크에서도 PageRank가 유일하게 정의된다.

하지만 Xing과 Ghorbani (2004)는 실제 웹페이지 간 링크들이 다른 중요도를 지니는 반면에 기존의 PageRank 알고리즘에서는 모든 링크가 동일한 weight를 가지고 있다는 한계점을 지적하고, edge의 weight를 고려한 weighted PageRank 알고리즘을 제안하였다. 본 논문에서는 weighted PageRank 알고리즘을 사용하였다.

3.1.2. Vertex centrality PageRank 알고리즘은 웹페이지의 중요도를 계산하기 위한 목적으로 만들어진 방법이며, 일반적으로 네트워크 분석에서는 이와 다른 다양한 중심도 지표를 사용한다. 주로 많이 사용되는 지표들에는 closeness, betweenness, eigenvector centrality가 있다 (Kolaczyk와 Csárdi, 2014). 이 개념을 설명하기 위해서는 특허 간의 간접적 인용 및 거리와 같은 개념을 먼저 이해해야 한다. 특허 i가 j를 인용하고, j가 k를 인용한 경우 특허 i가 k를 직접적으로 인용한 것은 아니지만 간접적으로 인용한다고 볼 수 있다. 이처럼 두 단계를 거친 인용 외에도 특허 i가 여러 인용 단계를 거쳐 k에 도달한 경우 이 둘의 관계를 간접적 인용이라고 한다. 그리고 특허 간 직접적, 간접적 인용 과정 중 최단 경로의 길이를 두 특허의 거리로 정의한다.

- Closeness centrality: 다른 특허들과의 거리가 가까울수록 그 특허가 중요하다는 개념을 사용한 것으로 짧은 인용 과정을 거쳐 다른 특허에 도달할수록 중요한 역할을 하는 특허라고 평가하는 지표다. 즉, closeness centrality는 다른 특허들과의 거리 합의 역수로 나타낸다.
- Betweenness centrality: 어떤 특허가 다른 두 특허 간의 더 많은 혹은 더 강한 간접적인 인용의 관계에 있을수록 그 특허가 중요하다는 개념을 사용한 지표다. 즉, 특허 i가 k를 인용하기 위해서는 반드시 j를 거쳐야 한다면 j가 두 특허를 연결시키는 역할을 한다고 할 수 있다. 이처럼 다른 특허들 사이에 위치하여 이들을 연결시켜주는 역할에 중점을 둔 측도라고 할 수 있다.
- Eigenvector centrality: 어떤 특허의 중요도는 그 특허를 인용한 특허들의 중요도에 따라 정해진다는 개념을 사용한 것으로, 적절한 선형식의 eigenvector solution을 구하는 방식으로 계산한다.

3.2. 네트워크 모형

네트워크 분석에서 다양한 모형을 사용할 수 있지만 본 논문에서는 특허 기술군의 군집화 분석을 위해 stochastic block model과 latent space model을 사용하였다. Stochastic block model을 통해 특허 기술군의 군집을 살펴본 후 이 결과와 latent space model 적합 결과를 연결지어 특허 기술군의 군집을 설명하고자 한다. 특허 인용 자료에서의 군집화 분석은 특허 간 인용관계를 설명하는 데 유용하게 사용될 수 있으며, 같은 군집으로 분류된 특허 기술군들의 특징과 유사점을 살펴보는 것 또한 큰 의미가 있다.

3.2.1. Stochastic block model Stochastic block model은 특허 기술군들이 Q개의 class로 분류되는 구조를 가지는 모형으로, class 분류에 따라 기술군 간 인용이 발생할 확률이 결정된다. 즉 기술군 간 인용이 발생할 확률은 오로지 class membership에만 의존한다. 구체적으로 i라는 특허 기술군이 class q에, j는 r에 속할 때 i가 j를 인용하는 경우 1, 인용하지 않는 경우 0값을 가지는 확률변수 Y_{ij}는 다음과 같은 베르누이 분포를 따른다고 가정한다 (Latouche 등, 2012).

$$Y_{ij}|(i \in q, j \in r) \sim \mathcal{B}(\pi_{qr}),$$

여기에서 $\boldsymbol{\pi}$는 class간 인용 발생 확률을 나타내는 $Q \times Q$ 행렬이다. 또한 특허 기술군 i가 각 class에 속할 확률 분포를 다음과 같은 벡터 $\boldsymbol{\alpha}$로 가정한다.

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q), \quad \text{where } \alpha_q = \Pr(i \in q) \quad \text{and } \sum_q \alpha_q = 1.$$

이 경우 $\boldsymbol{\pi}$와 $\boldsymbol{\alpha}$를 모두 추정하기 위해서 EM알고리즘을 이용한 variational approach 방법을 사용할 수 있다 (Daudin 등, 2008).

3.2.2. Latent space model 앞에서 설명한 stochastic block model의 경우 특허 기술군들의 class를 모형 적합에 사용하였다. 이는 우리가 관측할 수는 없지만 기술군 간에 인용이 발생할 확률, 즉 edge probability에 영향을 미친다고 여겨지는 잠재 변수를 모형에 포함시킨 것과 같다. 반면에 latent space model은 보다 일반적인 잠재 변수를 모형 적합에 사용한 경우다. Latent space model은 모든 특허 기술군을 d-차원 latent space 위의 점에 대응시킬 수 있다고 가정하며 그 위치를 잠재 변수의 형태로 모형에 포함시킨다. 또한 stochastic block model의 경우 인용의 여부만 반응변수로 고려하지만 latent space model의 경우 인용횟수를 반응변수로 고려할 수 있다. Latent space model은 다음과 같이 표현할 수 있으며 기술군 간의 인용, 즉 각 edge들은 서로 독립이라고 가정한다 (Krivitsky와 Handcock, 2008).

$$\begin{aligned}\Pr(Y = y|\beta, x, Z) &= \prod_{(i,j)\in\mathcal{Y}} \Pr(Y_{ij} = y_{ij}|\beta, x_{\cdot ij}, Z_i, Z_j) \\ &= \prod_{(i,j)\in\mathcal{Y}} f(y_{ij}|\mathrm{E}(Y_{ij}|\beta, x_{\cdot ij}, Z_i, Z_j)),\end{aligned}$$

여기서 $\{x_{kij}\}_{k=1}^p$는 기술군 i와 j 사이의 edge covariate를 나타내며, 이는 두 기술군 간 인용의 여러가지 특성을 나타내는 변수다. 또한 β는 edge covariate의 coefficient를 나타내며, Z_i, Z_j는 기술군 i와 j의 latent space에서의 위치를 나타내는 잠재 변수다. 인용 횟수 y의 조건부 분포에 따라 확률 분포 함수 f가 정해지는데, 베르누이 분포, 이항분포, 포아송 분포가 사용될 수 있다. y의 조건부 평균 $\mathrm{E}(Y_{ij}|\beta, x_{\cdot ij}, Z_i, Z_j)(= \mu_{ij})$은 다음과 같이 link function g를 통해 linear predictor function η_{ij}와 연결된다.

$$\eta_{ij} = g(\mu_{ij}) = \sum_{k=1}^{p} x_{kij}\beta_k - |Z_i - Z_j|.$$

위의 식에서 알 수 있듯이 latent space model은 두 특허 기술군의 latent space 거리가 인용이 발생할 확률에 영향을 미치는 구조를 가지는 모형이다. 모형의 추정에는 최대 우도 추정법 또는 MCMC sampling을 이용한 베이지안 방법을 사용한다 (Handcock 등, 2007). 이에 따라 모수의 점 추정에 최대 우도 추정치(MLE), 사후분포 평균(posterior mean), 사후분포 최빈수(posterior mode)를 사용할 수 있지만 이 세 가지 추정치는 많은 경우에 좋은 결과를 나타내지 못하기 때문에 Minimum Kullback-Leibler(MKL) 추정치를 사용한다 (Shortreed 등, 2006).

4. 분석 결과

4.1. 중심도 지표

앞에서 소개한 중심도 지표를 이용하여 크게 세 가지를 알아보고자 한다. 첫째, 특허 인용 네트워크에서 PageRank 알고리즘을 이용하여 주요 특허 기술군을 찾는다. 둘째, 네트워크 분석에서 일반적으로 쓰이

Table 4.1. Top 10 C3

	C3	PageRank	Contents
1	G06F	0.0350	Electric Digital Data Processing
2	H01L	0.0285	Semiconductor Devices; Electric Solid State Devices not Otherwise Provided for
3	A61B	0.0172	Diagnosis; Surgery; Identification
4	B32B	0.0157	Layered Products, i.e. Products Built-Up of Strata of Flat or Non-Flat
5	A61K	0.0156	Preparations for Medical, Dental, or Toilet Purposes
6	B65D	0.0143	Containers for Storage or Transport of Articles or Materials
7	G01N	0.0132	Investigating or Analysing Materials by Determining Their Chemical or Physical Properties
8	A61M	0.0125	Devices for Introducing Media into, or onto, the Body
9	B01D	0.0118	Separation
10	H04N	0.0116	Pictorial Communication

는 세 가지 centrality를 이용하여 주요 특허 기술군을 찾은 뒤 PageRank 알고리즘을 이용하여 얻은 결과와 비교한다. 셋째, 시대별로 주요 특허 기술군이 어떻게 변화하는지를 살펴본다. 여기에서는 시대별로 출원된 특허들 중에서 주요 특허 기술군을 찾기 위해 PageRank 알고리즘을 사용한다. 이와 같은 중심도 지표를 이용한 분석에는 R package 'igraph'를 사용하였다 (Csárdi와 Nepusz, 2006).

총 특허의 개수가 3,834,379개로 매우 많기 때문에 개별 특허의 중요도를 구하는 것보다 각 특허가 속한 C3 기술군의 중요도를 알아보는 것이 더 의미있는 결과를 얻을 수 있다고 생각되어 중심도 지표 분석에서는 C3 기술군을 vertex로 갖는 네트워크를 이용하였다. 이 경우 하나의 vertex는 하나의 C3 기술군을 나타내며 이 안에 여러 특허가 포함된다. 또한 두 개 vertex 사이의 edge 개수는 두 기술군 사이의 인용 횟수를 나타낸다.

먼저, PageRank 알고리즘을 특허 인용 네트워크에 적용할 때에는 weighted 네트워크를 이용하였다. 앞에서 설명한 네트워크에서 두 기술군 간의 인용 횟수를 edge의 weight로 설정한다. 이러한 weighted edge를 가지는 네트워크는 633개의 vertex와, 155,813개의 edge(loop, 즉 C3 기술군의 자기 인용 포함)로 구성된다. 이 네트워크에서 weighted PageRank 알고리즘으로 찾은 주요 C3 기술군 중 상위 10개 기술군이 다음 Table 4.1에 나열되어 있다. 여기에서 PageRank 점수가 클수록 중요한 기술군임을 나타낸다.

Table 4.1을 보면 G06F가 가장 높은 PageRank 점수를 나타내며 다음으로 H01L이 높은 점수를 나타내고 있는데, IT 기술의 발전으로 이와 관련된 G06, H01과 같은 기술군이 높은 중요도를 갖는 것으로 해석할 수 있다. 또한 의료기기와 관련된 A61B, A62K, A61M 등의 기술군이 주요 기술군으로 나타나고 있다.

다음으로 일반적인 네트워크 분석에서 사용하는 vertex centrality를 이용하여 주요 특허 기술군을 찾고 PageRank 알고리즘을 사용했을 때의 결과와 비교해 보았다. 앞서 설명한 세 가지 centrality의 개념을 살펴보면 eigenvector centrality가 PageRank 알고리즘과 가장 비슷한 개념을 사용하고 있다는 것을 알 수 있다. 주변의 이웃한 vertex의 중요도에 따라 자신의 중요도가 정해지며 이러한 순환적인 구조 때문에 eigenvector 해를 구하는 방식으로 중요도를 계산한다는 점이 유사하다. 실제로 상위 10개 주요 C3 기술군을 찾을 때, PageRank 알고리즘을 이용한 것과 세 가지 centrality를 이용한 것을 비교하면 Table 4.2에서 볼 수 있듯이 eigenvector centrality가 PageRank 알고리즘과 가장 비슷한 결과를 나타낸다.

특허 인용 네트워크의 경우 directed graph이기 때문에 두 특허 기술군 사이의 거리를 잴 때, 기준이 되

Table 4.2. Comparison between PageRank and vertex centrality

	Closeness (in)	Closeness (out)	Betweenness	Eigenvector	PageRank
1	C13C	B22D	G03F	G06F	G06F
2	B60H	B66F	G03C	H04L	H01L
3	B41F	G03C	A01N	H04N	A61B
4	D06F	D05B	C12P	G06K	B32B
5	A01C	B60J	B29D	H04M	A61K
6	A22C	C03B	D05B	G11C	B65D
7	B07B	B29D	B30B	G09G	G01N
8	C06B	C08J	A47L	H04J	A61M
9	H03F	F42B	G01S	H01L	B01D
10	H02B	B31F	A61K	H04Q	H04N

Table 4.3. Top 10 C3

	1985–1989	1990–1994	1995–1999	2000–2004	2005–2009	2010–2012
1	B32B	G06F	G06F	G06F	G06F	G06F
2	G06F	B65D	H01L	H01L	H01L	H01L
3	B65D	H01L	A61K	A61K	G02B	A61K
4	H04N	A61K	A61B	A61B	H04L	A61B
5	H01L	B32B	B32B	B32B	A61K	B01D
6	A61K	B01D	B65D	G02B	G01N	B32B
7	G01N	G02B	H04N	B01D	G06K	H04B
8	B01D	G01N	B01D	G01N	H04N	G06K
9	G02B	A61B	G11B	B65D	H01R	G01N
10	G11B	H04N	G01N	A61F	H04B	G02B

는 특허가 간접적 인용이 되는 특허인지 간접적 인용을 하는 특허인지에 따라 두 가지 타입의 closeness centrality가 정의될 수 있다. Table 4.2에서 closeness (in)은 다른 특허들이 기준이 되는 특허를 간접적으로 인용하는 과정의 길이를 재는 경우이며, closeness (out)은 기준이 되는 특허가 다른 특허들을 간접적으로 인용하는 과정의 길이를 재는 경우를 나타낸다. 이 두 경우와 betweenness centrality 모두 PageRank 알고리즘과는 매우 다른 결과를 보이고 있는 반면에 eigenvector centrality가 PageRank 알고리즘과 개념은 물론 분석 결과가 가장 비슷하다는 것을 알 수 있다.

특허가 중요한지 아닌지를 판단하기 위해서는 그 특허가 얼마나 많은 특허로부터 인용되었는지, 얼마나 중요한 특허로부터 인용되었는지를 고려해야 한다. 하지만 closeness centrality는 다른 특허들로부터 가까운 거리에 있는 특허를 중요하다고 여기며, betweenness centrality는 특허와 특허를 연결하는 path에 위치하여 이들을 연결시켜주는 특허를 중요하다고 여긴다. 이는 일반적으로 생각하는 중요한 특허가 지니는 특성과는 관련이 없다. 따라서 특허 인용 네트워크에서 주요 기술군을 파악하기 위해서는 PageRank 알고리즘이나 eigenvector centrality를 사용하는 것이 적절하다고 볼 수 있다.

마지막으로 시간이 흐르면서 그 시대를 대표하는 기술군이 어떻게 변화하는지 알아보고자 한다. 시대별 분석을 위해 1985년부터 2012년까지의 자료를 6개의 시기로 나누어 주요 기술군을 살펴보았다. 5년 간격으로 1985–1989년, 1990–1994년, 1995–1999년, 2000–2004년, 2005–2009년, 2010–2012년의 6개의 시기로 구분하였다. 각 기간에서 출원된 특허들 간의 인용 자료를 바탕으로 PageRank 알고리즘을 사용하여 찾은 상위 10개 주요 C3 기술군을 Table 4.3에 표시하였다.

90년대 후반부터는 G06F와 H01L과 같은 IT 관련 기술군이 계속해서 상위를 차지하고 있으며 과거에

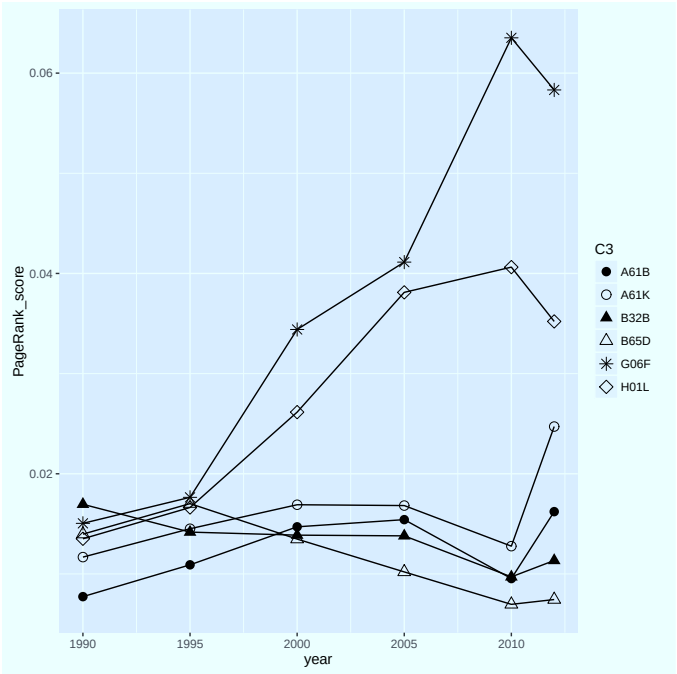


Figure 4.1. Changing PageRank score.

비해 A61K나 A61B와 같은 의료기기 관련 기술군이 중요한 기술군으로 뽑히고 있다. 반대로 B32B나 B65D와 같은 운수 관련 기술군은 순위가 점차 낮아지고 있음을 알 수 있다. 이와 같은 사실은 Figure 4.1에서도 확인할 수 있는데 이는 A61B, A61K, B32B, B65D, G06F, H01L의 6가지 기술군의 PageRank 점수가 시대별로 어떻게 변화하는지를 나타낸 그림이다. G06F와 H01L 기술군이 단순히 주요 기술군 순위의 상위권에 있는 것이 아니라 PageRank 점수가 90년대 초반에 비해 대폭 상승한 것으로 보아 중요도가 점점 높아지는 것을 알 수 있다. 그리고 A61B와 A61K 기술군은 PageRank 점수가 조금씩 증가하는 추세이며, B32B와 B65D 기술군은 반대 양상을 띠고 있다.

4.2. 네트워크 모형

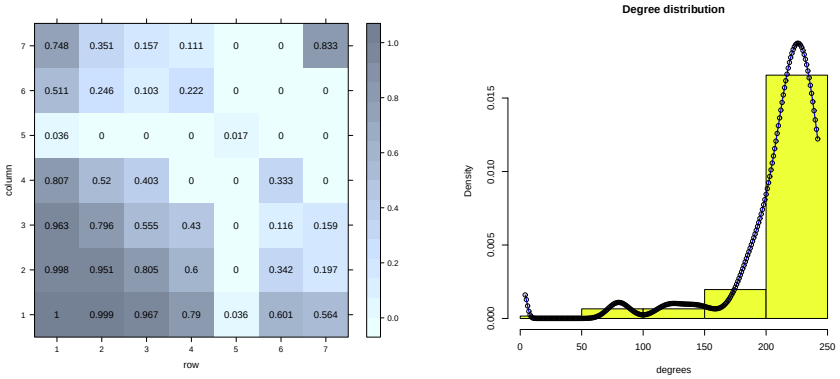
4.2.1. Stochastic block model Stochastic block model 분석에서는 R package 'mixer'를 사용하였다 (Daudin 등, 2008). 특허 인용 네트워크에 Stochastic block model을 적합시킬 때에는 C2 기술군을 vertex로 갖는 네트워크를 사용하였다. 이 경우는 122개의 vertex와, 13,052개의 edge(loop, 즉 C2 기술군의 자기 인용 포함)로 구성된 네트워크이다.

Variational approach 추정 방법을 사용할 경우 class의 개수 Q는 Integration Classification Likelihood(ICL)을 기준으로 정해진다. 이 기준을 사용할 경우 특허 인용 네트워크에 속한 vertex인 C2 기술군은 7개의 class로 분류된다. Table 4.4는 각 class에 어떤 vertex가 속하는지를 나타낸다.

C2 기술군이 분류된 것을 살펴보면, 첫 번째와 두 번째 class는 각각 57개, 38개의 기술군으로 이루어져 있으며, 나머지 class들은 비교적 적은 수의 기술군으로 이루어져 있다. 가장 주목할 만한 class는 첫 번째 class인데 주요 기술군 대부분이 여기에 속해 있다. 앞서 PageRank 알고리즘을 이용하여 찾은 주요

Table 4.4. Classification of vertices

Class	Vertex
Class1 (57)	A01, A23, A47, A61, A63, B01, B05, B08, B21, B22, B23, B24, B25, B26, B28, B29, B32, B41, B60, B63, B64, B65, B67, C03, C04, C07, C08, C09, C10, C12, C23, C25, D06, E04, E21, F01, F02, F04, F16, F24, F25, F26, F28, G01, G02, G03, G05, G06, G08, G09, G11, G21, H01, H02, H03, H04, H05
Class2 (38)	A21, A24, A41, A45, A62, B02, B03, B07, B27, B30, B31, B42, B44, B61, B62, B66, C01, C02, C11, C22, D01, D02, D03, D04, D21, E01, E02, E03, E05, E06, F15, F17, F21, F23, F27, F41, F42, G10
Class3 (17)	A22, A42, A43, A44, A46, B04, B06, B09, B43, C06, C21, C30, D05, F03, F22, G04, G07
Class4 (3)	B68, C05, G12
Class5 (1)	A99
Class6 (3)	C13, C14, D07
Class7 (3)	B81, B82, C40



(a) Edge probability (b) Goodness-of-fit

Figure 4.2. Result of stochastic block model.

C3 기술군을 나타낸 Table 4.1을 살펴보면 이들이 속한 C2 기술군 모두 첫 번째 class에 속하는 것을 확인할 수 있다. Table 4.1에는 상위 10개 기술군만 표시했지만 상위 50개까지 추가로 확인한 결과 이들 또한 모두 첫 번째 class로 분류되는 것을 알 수 있었다. 다음으로 살펴볼 class는 다섯 번째 class로 A99라는 기술군 하나로 이루어져 있다. 이 기술군은 A(생활필수품) section 안에서 다루어지지 않은 '기타' 항목을 나타낸다. A99 기술군의 in degree와 out degree는 모두 2이며, 이 분석에서 다루는 약 400만 개의 특허 중 하나의 특허만 이 기술군에 포함되어 있다. 이와 같은 사실로 미루어 보았을 때, 이 기술군이 다른 기술군을 인용하거나 다른 기술군이 이 기술군을 인용하는 경우는 극소수이며 포함된 특허도 거의 없다고 볼 수 있다. 즉 전체 네트워크에서 중요한 역할을 하지 않는 vertex라고 볼 수 있다.

Figure 4.2의 (a)는 class간 연결 확률 행렬인 $\boldsymbol{\pi}$를 나타내는데 이는 class들의 특징과 연결지어 해석할 수 있다. 첫 번째 class는 PageRank 알고리즘을 이용해서 찾은 주요 기술군들이 주로 포함된 그룹이다. 여기서 주요 기술군이라는 것은 다른 중요한 기술군으로부터 인용을 받았거나 다른 기술군들과의 인용 관계가 활발하다는 것을 뜻한다. 그렇기 때문에 다른 기술군과의 연결 확률이 전반적으로 높다. 반면

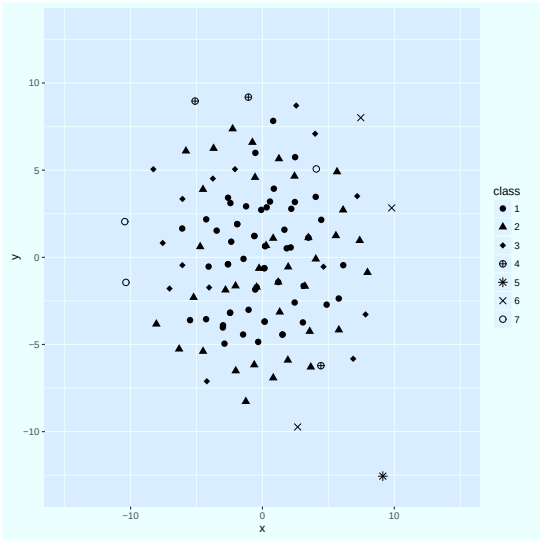


Figure 4.3. Latent space position.

다섯 번째 class는 전체 네트워크에서 크게 중요한 역할을 하지 않는 기술군으로 이루어져 있으며 다른 class들과 인용관계를 갖는 경우가 거의 없기 때문에 다른 기술군과의 연결 확률이 대부분 0으로 추정되었다.

마지막으로 Figure 4.2의 (b)는 적합된 모형이 실제로 관측된 네트워크를 얼마나 잘 설명하는지 알아보기 위한 degree의 적합도 검정을 보여준다. 이 그림에서 히스토그램과 곡선은 각각 관측된 네트워크와 적합된 모형의 degree distribution을 나타낸다. 히스토그램과 곡선이 비슷한 형태를 지니고 있는 것으로 보아 적합된 stochastic block model이 특허 인용 네트워크를 잘 설명하고 있다고 할 수 있다.

4.2.2. Latent space model Latent space model 분석에서는 R package 'latentnet'를 사용하였다 (Krivitsky 와 Handcock, 2008). 이 모형을 적합시킬 때에도 마찬가지로 C2 기술군을 vertex로 갖는 네트워크를 사용하였다. C2 기술군 간의 인용횟수를 edge의 weight로 가지는 네트워크이기 때문에 y의 조건부 분포 f는 포아송 분포, link function g는 로그 함수로 가정하고 분석을 진행하였다. 이 분석에서는 edge covariate를 사용하지 않았으며 2차원 latent space를 가정하였다. 적합 결과 y의 조건부 분포는 다음과 같고, 추정한 latent space position은 Figure 4.3과 같다. 모형 추정에는 MCMC samping을 이용한 베이지안 방법을 사용하여 다음과 같은 모형을 구하였다.

$$\log E(Y_{ij}|Z_i, Z_j) = 11.390 - |Z_i - Z_j|.$$

Figure 4.3의 vertex는 앞서 stochastic block model 적합 결과 분류된 class에 따라 다른 모양의 점으로 나타냈다. 첫 번째, 두 번째, 세 번째 class가 중심부에 모여있고 네 번째, 여섯 번째, 일곱 번째 class은 각 3개의 vertex를 가지면서 상대적으로 가장자리에 위치하고 있다. 다섯 번째 class는 A99 기술군만 포함하고 있으며 다른 vertex들과 동떨어진 자리에 위치하고 있음을 확인할 수 있다.

Latent space position은 class간 연결 확률 행렬과 연결지어 살펴볼 수 있다. 중심부에 모여 있는 첫 번째, 두 번째, 세 번째 class는 2차원의 latent space에서 가까운 거리에 위치하고 있기 때문에 연결의 빈도가 높으며 서로 인용하는 경우가 많다고 가정할 수 있다. 하지만 상대적으로 가장자리에 위치

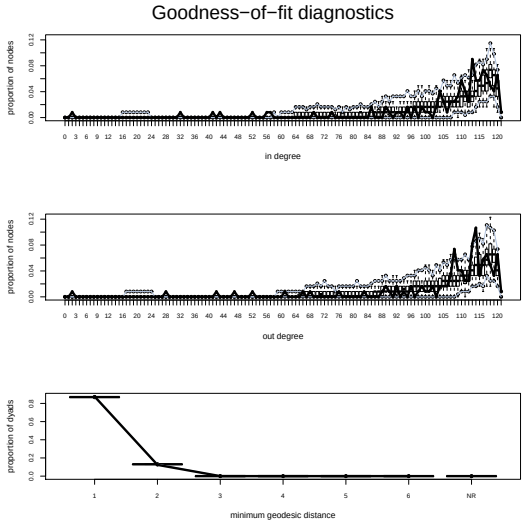


Figure 4.4. Goodness of fit test.

한 class들은 중심부에 모여있는 class와 거리가 멀기 때문에 연결의 빈도가 낮으며 인용횟수도 적으리라 생각된다. 이와 같은 사실은 Figure 4.2의 (a)가 나타내는 stochastic block model의 class 간 연결확률 행렬에서도 확인할 수 있다.

다음으로 적합된 모형이 얼마나 적절한지 평가하기 위해 실제로 관측된 네트워크가 적합된 모형에서 시뮬레이션을 통해 생성된 네트워크와 얼마나 비슷한지 비교해볼 수 있다. 이 두 네트워크를 비교하는 데에는 in degree, out degree와 minimum geodesic distance와 같은 통계량이 사용된다. 여기에서 minimum geodesic distance는 두 vertex 사이의 최단 경로의 길이를 뜻한다. Figure 4.4를 통해 이 세 가지 통계량의 분포가 두 네트워크에서 어떻게 다르게 나타나는지 확인할 수 있다. 실제로 관측된 네트워크의 통계량 분포를 검은 실선으로, 시뮬레이션을 통해 생성된 네트워크의 통계량 분포를 상자그림으로 나타냈다. 여기에서는 100번의 시뮬레이션을 시행하였다. In degree와 out degree의 경우 두 분포가 정확하게 같지는 않지만 비슷한 추세를 보이고 있는 것을 보아 적합된 모형이 관측된 네트워크의 통계량을 비교적 잘 나타내는 것으로 볼 수 있다. Minimum geodesic distance의 경우 두 분포가 거의 일치한다. 이 그림에서 확인할 수 있듯이 두 vertex간 minimum geodesic distance가 1인 경우가 대부분이다. 이것은 대부분의 vertex가 서로 직접적으로 연결되어있으며 네트워크의 density가 높다는 것을 의미하기도 한다. 적합된 모형에서 생성된 네트워크의 minimum geodesic distance는 실제 네트워크를 거의 완벽하게 재현하며 세 가지 통계량을 통해 적합도를 검정한 결과 적합된 latent space model이 관측된 네트워크를 전반적으로 잘 설명하고 있는 것으로 해석할 수 있다.

Latent space model은 두 특허 기술군 간 latent space에서의 거리와 인용횟수가 반비례한다고 설명하고 있으며, 이와 같은 사실은 stochastic block model 적합 결과로 얻은 그룹 분류와 연결지어 파악할 수 있었다. Latent space의 중심부에 위치한 첫 번째, 두 번째, 세 번째 class는 서로 인용하는 빈도가 높고, 가장자리에 위치한 네 번째, 여섯 번째, 일곱 번째 class는 다른 class들과의 인용관계 빈도가 낮다. 마지막으로 다른 기술군들과 멀리 떨어져있는 다섯 번째 class는 다른 기술군들과의 인용횟수가 현저하게 낮게 나타난다. 이처럼 latent space model을 적합하여 얻은 특허 기술군 간 인용관계는 stochastic block model의 결과와 비슷한 맥락에서 해석할 수 있다.

5. 결론

본 연구에서는 피인용 횟수가 많은 것 뿐 아니라 얼마나 중요한 특허로부터 인용되었는지를 고려하기 위해 PageRank 알고리즘을 사용하여 기술군의 중요도를 계산하였다. 일반적으로 중요하다고 생각되는 특허가 가지는 특징을 반영하기 위해서는 네트워크 분석에서 주로 사용되는 중심도 지표보다 PageRank 알고리즘을 사용하는 것이 적절하다고 할 수 있다. 그 결과 IT와 의료기기 관련 기술군이 새로운 트렌드를 주도하는 것으로 나타났다. 또한 본 연구에서는 특허 인용 네트워크에 두 가지 네트워크 모형을 적합하여 특허 기술군의 군집화를 분석하였다. 그 결과 기술군의 그룹이 나뉘는 데에는 각 기술군의 중요도가 영향을 미친다는 사실을 확인하였으며, 군집의 분류에 따라 특허 기술군 간의 인용 관계를 설명할 수 있었다. 다음으로 특허 인용 네트워크를 2차원 latent space에 나타냄으로써 그룹의 특징과 그룹 간의 관계를 다시 한 번 파악할 수 있었다. 본 연구는 1985년부터 2012년 사이에 미국 특허청에 등록된 약 400만 개의 모든 특허를 사용하여 전반적인 특허 기술의 동향을 살펴보았지만, 목적에 따라 특정 기술군과 관련된 특허만 추출하여 같은 방법으로 분석을 진행할 수 있을 것이다.

References

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, **30**, 107–117.

Csárdi, G. and Nepusz, T. (2006). The igraph software package for complex network research, *Interjornal*, Complex Systems, 1695.

Daudin, J. J., Picard, F., and Robin, S. (2008). A mixture model for random graphs, *Statistics and Computing*, **18**, 173–183.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170**, 301–354.

Kolaczyk, E. D. and Csárdi, G. (2014). *Statistical Analysis of Network Data with R*, Springer, New York.

Krivitsky, P. N. and Handcock, M. (2008). Fitting position latent cluster models for social networks with latentnet, *Journal of Statistical Software*, **24**(i05).

Latouche, P., Birmele, E., and Ambroise, C. (2012). Variational Bayesian inference and complexity control for stochastic block models, *Statistical Modelling*, **12**, 93–115.

Li, X., Lin, Y., Chen, H., and Roco, M. C. (2007). Worldwide nanotechnology development: a comparative study of USPTO, EPO, and JPO patents (1976–2004), *Journal of Nanoparticle Research*, **9**, 977–1002.

Oh, S., Lei, Z., and Yen, J. (2012). Evaluating and ranking patents using weighted citations. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 281–284.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: bringing order to the web, *Stanford InfoLab*.

Shortreed, S., Handcock, M. S., and Hoff, P. (2006). Positional estimation within a latent space model for networks, *Methodology*, **2**, 24–33.

Xing, W. and Ghorbani, A. (2004). Weighted PageRank algorithm. In *Proceedings of the Second Annual Conference on Communication Networks and Services Research*, 305–314.

Yoo, S.-H., Lee, Y.-H., and Won, D.-K. (2007). A study on the measurement of technological impact using citation analysis of patent information, *Journal of Korea Technology Innovation Society*, **10**, 687–705.

특허 인용 네트워크 분석

이민정[a] · 김용대[b] · 장원철[b,1]

[a]네이버, [b]서울대학교 통계학과

(2016년 2월 11일 접수, 2016년 4월 14일 수정, 2016년 4월 25일 채택)

요 약

과학 기술의 발전은 사회를 급격하게 변화시켜 왔다. 특허 자료 분석은 현대 과학 기술의 흐름을 이해하고 미래 유망기술을 예측할 수 있게 한다. 본 연구에서는 기술의 동향을 파악하고자 1985년과 2012년 사이에 미국 특허청에 등록된 특허를 중심으로 특허 인용 네트워크를 분석한다. 주요 기술군을 파악하기 위해 PageRank 알고리즘 외에 다양한 중심성 지표를 이용하고, 통계적 네트워크 모형을 통해 유사한 기술들의 군집을 찾아내고자 한다.

주요용어: 특허 인용, 유망 기술, PageRank 알고리즘, stochastic block model, latent space model

이 논문은 2014년도 SNU Brian Fusion Program 지원과 2013년도 정부(교육부)와 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단 지원을 받아 수행된 기초연구사업임 (No. 2013R1A1A2010065, No. 2014R1A4A1007895).

[1]교신저자: (08826) 서울특별시 관악구 관악로1, 서울대학교 통계학과. E-mail: wcjang@snu.ac.kr