JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 8, pp. 143-149, August 2022
https://doi.org/10.9708/jksci.2022.27.08.143

Patent Keyword Analysis using Gamma Regression Model and Visualization

Sunghae Jun*

*Professor, Dept. of Big Data and Statistics, Cheongju University, Cheongju, Korea

[Abstract]

Since patent documents contain detailed results of research and development technologies, many studies on various patent analysis methods for effective technology analysis have been conducted. In particular, research on quantitative patent analysis by statistics and machine learning algorithms has been actively conducted recently. The most used patent data in quantitative patent analysis is technology keywords. Most of the existing methods for analyzing the keyword data were models based on the Gaussian probability distribution with random variable on real space from negative infinity to positive infinity. In this paper, we propose a model using gamma probability distribution to analyze the frequency data of patent keywords that can theoretically have values from zero to positive infinity. In addition, in order to determine the regression equation of the gamma-based regression model, two-mode network is constructed to visualize the technological association between keywords. Practical patent data is collected and analyzed for performance evaluation between the proposed method and the existing Gaussian-based analysis models.

▸ **Key words**: Patent Keyword, Generalized Linear Model, Gaussian Distribution, Gamma Distribution, Patent Analysis

[요 약]

특허문서는 연구 개발된 기술에 대한 상세한 결과를 포함하고 있기 때문에 효과적인 기술분석을 위한 다양한 특허분석 방법에 대한 연구가 진행되고 있다. 특히 통계학과 머신러닝 알고리즘에 의한 정량적인 특허분석에 대한 연구가 최근 활발하게 이루어지고 있다. 정량적 특허분석에서 가장 많이 사용되는 특허 데이터는 기술 키워드이다. 기술 키워드 데이터를 분석하는 기존의 방법은 대부분 음의 무한대부터 양의 무한대까지 실수 공간 전체를 확률변수의 값으로 갖는 가우시안 확률분포에 기반한 모형이었다. 본 논문에서는 이론적으로 0부터 양의 무한대까지의 값을 갖는 특허 키워드의 빈도 데이터를 분석하기 위하여 감마 확률분포를 활용한 모형을 제안한다. 또한 감마 회귀모형의 회귀방정식을 결정하기 위하여 키워드 간의 기술 연관성을 시각화하는 2-모드 네트워크를 구축한다. 제안 방법과 기존의 가우시안 기반의 분석모형 간의 성능평가를 위하여 실제 특허 데이터를 수집하여 분석한다.

▸ **주제어**: 특허키워드, 일반화선형모형, 가우시안분포, 감마분포, 특허분석

• First Author: Sunghae Jun, Corresponding Author: Sunghae Jun
*Sunghae Jun (shjun@cju.ac.kr), Dept. of Big Data and Statistics, Cheongju University
• Received: 2022. 07. 08, Revised: 2022. 08. 16, Accepted: 2022. 08. 17.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

과학기술이 발전할수록 특허의 중요성은 더욱 증가한다. 연구 개발된 기술 결과는 특허 시스템에 출원, 등록되어 자신이 개발한 기술에 대한 배타적인 권리를 일정 기간 확보하려고 한다. 특허문서에는 개발된 기술에 대한 자세한 설명이 포함되기 때문에 기술에 대한 정량적 분석을 위한 가장 가치 있는 데이터는 특허문서이다. 특허문서는 제목과 요약 뿐만 아니라 청구항(claims), 국제 특허분류 코드, 인용정보, 발명인, 출원날짜, 그림 등 개발된 기술에 대한 방대한 결과를 포함하고 있다 [1]. 따라서 우리는 특허분석을 통하여 개발된 기술에 대한 정량적 분석이 가능하고 필요에 따라 미래기술에 대한 예측까지 수행할 수 있게 되었다 [2,3]. 지금까지 특허분석을 위한 많은 연구가 다양한 분야에서 수행되어 왔다 [3-8]. 특히 특허문서로부터 추출된 기술 키워드 간의 연관성을 분석하는 키워드 분석에 대한 연구가 최근 활발하게 진행되고 있다 [2-8]. 대부분의 연구들은 통계학과 머신러닝의 일반화 선형모형(generalized linear model, GLM)에 기반하고 있다 [3,7-11]. 특허 키워드분석에서 사용되는 GLM의 연결함수(link function)는 대부분 가우시안 분포(Gaussian distribution)를 사용하고 있다. 가우시안 분포는 음의 무한대부터 양의 무한대까지 실수공간 전체를 확률변수의 값으로 정의하지만 특허 키워드의 데이터는 0부터 양의 무한대까지 값을 갖는다. 따라서 가우시안 분포에 기반한 GLM은 특허 키워드분석에서 한계를 보인다.

본 연구에서는 특허 키워드 데이터의 효율적인 분석을 위하여 감마분포(gamma distribution)를 연결함수로 갖는 GLM을 사용한다. 또한 키워드 간의 연관관계를 시각화하기 위하여 그래프 이론(graph theory)에 기반한 2-모드 네트워크(two-mode network)를 구축한다. 2-모드 네트워크 시각화 결과로부터 회귀식(regression equation)을 결정하고 회귀식의 모수를 감마분포 기반의 GLM을 이용하여 추정하는 특허 키워드분석 방법을 제안한다. 본 논문의 제안방법과 기존의 가우시안 기반의 GLM과의 성능비교를 위하여 모형의 설명력을 측정하는 통계적 지표를 이용한다.

본 논문의 2장에서는 기존의 기술 분석 방법에 대하여 관련된 연구를 알아본다. 제안하는 기술 분석 방법은 3장에서 다루고, 4장에서는 제안방법과 기존방법의 성능비교를 위하여 실제 특허문서를 수집하고 분석한다. 마지막 장에서는 본 연구에 대한 결론 및 한계점과 향후 연구과제를 제시한다.

II. Related Works

기술에 대한 연관관계를 파악하고 기술예측을 위하여 특허 분석이 수행된다 [1]. 분석에 사용되는 특허문서는 전세계 특허 데이터베이스로부터 검색한다 [12]. 통계학과 머신러닝에서 제공하는 분석 알고리즘을 이용하여 수집된 특허문서를 분석하기 위하여 우리는 특허문서에 대한 전처리(preprocessing)를 수행한다. 즉, 텍스트 마이닝의 자연어 처리기법을 이용하여 특허문서로부터 키워드를 추출하여 최종적으로 특허-키워드 행렬(patent-keyword matrix)을 구축한다 [13,14]. 이 행렬의 행과 열은 각각 특허문서와 키워드를 나타낸다. 행렬의 원소는 특허문서에 포함된 키워드의 빈도수이다. 정량적 특허분석 과정에서 키워드는 변수(variables)로 사용된다. 예를 들어 선형회귀모형(linaer regression model)에서 키워드는 반응변수와 설명변수로 사용된다,

일반적으로 특허-키워드 행렬의 빈도 데이터를 분석하기 위하여 선형회귀모형을 포함한 GLM을 사용한다 [3,11,15]. GLM에서 반응변수의 확률함수는 지수족(exponential family)이고 평균 모수는 설명변수의 선형결합(linear combination)으로 표현된다 [16]. GLM은 연결함수를 통하여 주어진 데이터에 가장 적합한 확률분포를 사용하여 모형을 구축한다. 특허분석을 위한 기존의 GLM에서는 식 (1)의 가우시안 확률분포를 사용하였다 [16-18].

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty \quad (1)$$

여기서 μ와 σ는 각각 가우시안 분포를 따르는 확률변수 X의 평균과 표준편차를 나타낸다. X는 음의 무한대부터 양의 무한대까지 모든 실수 값을 가질 수 있다. 그러나 0부터 매우 큰 양의 값까지 가질 수 있는 특허 키워드의 빈도 데이터의 분석에서 가우시안 연결함수를 갖는 GLM은 모형의 설명력에서 한계를 보인다. 이와 같은 문제를 해결하기 위하여 본 연구에서는 확률변수가 0보다 큰 값을 갖는 감마분포 기반의 GLM을 사용하는 특허분석 방법을 제안한다.

III. Proposed Method

본 논문에서 우리는 가우시안 확률분포 기반의 GLM에 비해 성능이 향상된 특허 키워드 분석모형을 구축하기 위

하여 2-모드 네트워크와 감마분포를 갖는 GLM을 결합한 특허분석 방법을 제안한다. 목표 기술과 관련되어 검색된 특허문서는 분석을 위하여 정형화된 데이터 구조로 전처리 되어야 한다. 본 논문에서는 텍스트 마이닝 기법을 사용한 전처리 과정을 수행한다. 먼저 검색된 특허문서로부터 제목과 요약 데이터를 추출하여 텍스트 데이터베이스를 구축하고 이를 바탕으로 코퍼스(corpus)를 생성한다. 생성된 코퍼스로부터 키워드를 추출하여 특허문서와 키워드로 이루어진 정형화된 데이터를 구축하여 최종적으로 표 1과 같은 특허-키워드 행렬을 얻는다.

Table 1. Patent document-Keyword matrix

	K_1	K_2	⋯	K_m
P_1	$freq._{11}$	$freq._{12}$	⋯	$freq._{1m}$
P_2	$freq._{21}$	$freq._{22}$	⋯	$freq._{2m}$
⋮	⋮	⋮	⋱	⋮
P_n	$freq._{n1}$	$freq._{n2}$	⋯	$freq._{nm}$

이 행렬의 행은 n개의 특허문서($P_1, P_2, ..., P_n$)를 나타내고 열은 m개의 키워드($K_1, K_2, ..., K_m$)를 나타낸다. 이 행렬의 각 원소는 각 특허문서에 포함된 특정 키워드의 빈도 값($freq._{11}, freq._{12}, ..., freq._{nm}$)이다. 빈도 값이 가질 수 있는 값의 범위는 0부터 양의 무한대까지이다. 기술 분석을 위하여 m개의 키워드 중에서 반응변수로 사용될 q개의 키워드($Y_1, Y_2, ..., Y_q$)와 설명변수로 사용될 p개의 키워드($X_1, X_2, ..., X_p$)를 선정한다. 2-모드 네트워크 시각화를 위하여 표 2와 같이 설명변수와 반응변수 간의 상관계수 행렬(correlation coefficient matrix)를 구한다.

Table 2. Correlation matrix of response and explanatory variables

	Y_1	Y_2	⋯	Y_q
X_1	$cor(x_1,y_1)$	$cor(x_1,y_2)$	⋯	$cor(x_1,y_q)$
X_2	$cor(x_2,y_1)$	$cor(x_2,y_2)$	⋯	$cor(x_2,y_q)$
⋮	⋮	⋮	⋱	⋮
X_p	$cor(x_p,y_1)$	$cor(x_p,y_2)$	⋯	$cor(x_p,y_q)$

본 논문에서 2-모드 네트워크는 p개의 행과 q개의 열 노드 간의 연결관계를 이용한 사회 네트워크(social networks) 구조를 갖는다. 즉, p개의 설명 키워드와 q개의 반응 키워드 간의 기술적 연관성을 시각화한다. 2-모드 네트워크를 생성하기 위하여 먼저 식 (2)와 같은 입력행렬(input matrix)을 만든다 [19,20].

$$M = \begin{pmatrix} 0_p & C \\ C^t & 0_q \end{pmatrix} \tag{2}$$

여기서 C는 표 2의 $(p \times q)$ 상관계수 행렬로 이루어진 연관행렬(affiliation matrix)이다. 0_n과 0_m은 각각 $(n \times n)$과 $(m \times m)$인 영(zero) 행렬이다. 최종적으로 행렬 M을 시각화 한 결과를 이용하여 감마분포 기반 GLM에 사용될 회귀식이 결정된다.

GLM은 반응변수, 선형예측자(linear predictor) 그리고 연결함수의 3가지 구성요소로 이루어진다. 서로 독립인 반응변수 $Y_1, Y_2, ..., Y_q$는 지수족에 속하는 분포를 따른다. 본 논문에서는 감마분포를 사용한다. 즉, 감마분포를 따르는 확률변수 Y는 식 (3)의 확률밀도함수를 갖는다 [21].

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} \exp\left(-\frac{y}{\beta}\right),\ 0 < y < \infty \tag{3}$$

확률변수 Y가 가질 수 있는 값은 0보다 큰 양의 실수이다. $X_1, X_2, ..., X_p$의 선형결합 $\sum_{j=0}^{p} X_j\beta_j$, $(X_0 = 1)$을 선형예측자라고 한다. Y_i의 평균을 μ_i라고 할 때 식 (4)의 관계를 갖는 $g(\cdot)$는 연결함수가 된다.

$$g(\mu_i) = \sum_{j=0}^{p} X_j\beta_j \tag{4}$$

본 논문에서는 특허 키워드의 빈도가 이론적으로 0부터 양의 무한대까지 산포가 크기 때문에 연결함수에 로그를 취하여 사용한다. 다음은 본 논문에서는 제안하는 특허 키워드 분석 절차를 나타낸다.

(Step 1) 특허문서의 수집 및 전처리
(1.1) 목표기술과 관련된 특허문서 검색
(1.2) 텍스트마이닝 기법을 이용한 특허 데이터의 전처리
(1.3) 특허-키워드 행렬 구축
(1.4) 반응변수와 설명변수로 사용될 기술 키워드 선정

(Step 2) 2-모드 네트워크 시각화
(2.1) 설명-반응 키워드 행렬 구축
(2.2) 설명-반응 키워드 간 상관계수 계산
(2.3) 영행렬과 연관행렬을 이용한 입력행렬 생성

(2.4) 2-모드 네트워크 시각화

(Step 3) 감마분포 기반 GLM 수행
(3.1) 반응 키워드의 분포를 감마 확률분포로 지정
(3.2) 설명 키워드에 대한 선형 예측자 정의
(3.3) 로그 연결함수를 이용하여 최종 GLM 수행
(3.4) 모형평가

본 논문에서 최종적으로 구축된 모형의 성능평가를 위하여 아카이케 정보기준(Akaike information criterion, AIC), 베이지안 정보기준(Bayesian information criterion, BIC) 그리고 로그우도(log likelihood)를 사용한다. 먼저 AIC는 식 (5)와 같이 정의된다 [17,22].

$$AIC = -2\log L + 2k \tag{5}$$

여기서 $\log L$은 로그우도이고 k는 추정해야 할 모수의 수를 나타낸다. AIC 값이 작을수록 모형의 성능은 우수하게 된다. AIC와 함께 모형의 성능평가에 주로 사용되는 BIC는 식 (6)과 같이 정의된다 [17,22].

$$BIC = -2\log L + k\log N \tag{6}$$

여기서 N은 데이터의 크기이다. 즉, 주어진 데이터가 큰 경우 BIC는 AIC에 비해 각 모수에 더 큰 패널티(penalty)를 갖게 한다. BIC도 AIC와 마찬가지로 작은 값을 가질수록 모형의 성능이 우수하게 된다. 본 논문에서 사용되는 세 번째 성능평가 측도는 식 (7)에서 정의되는 로그우도이다 [17,23].

$$\log L = \sum_{i}^{N} \log p(x_i \mid \beta) \tag{7}$$

로그우도는 주어진 데이터에 대한 모수의 확률함수에 로그를 취한 값의 합을 나타낸다. 즉, 로그우도는 모수에 대한 가능도를 나타내기 때문에 이 값이 클수록 모형의 성능이 우수하게 된다. 실제 특허 데이터를 이용한 본 논문의 실험에서 비교 모형 간의 성능평가를 위하여 우리는 AIC, BIC 그리고 로그우도 값을 사용한다.

IV. Experimental Results

기후변화 등으로 인하여 재난(disaster)이 지속적으로 증가하고 있고 이에 대처할 수 있는 기술개발이 절실한 시점이다. 따라서 본 연구에서는 제안 방법의 성능평가를 위하여 재난 대처 기술에 대한 특허문서를 특허 데이터베이스로부터 검색하였다 [24]. 최종적으로 선택된 재난 기술의 유효한 특허문서는 16,875건이었다. 대표적인 데이터 분석언어인 R과 R에서 제공하는 텍스트 마이닝 패키지를 사용하여 특허-키워드 행렬을 구축하였다 [13,14,25]. 표 3은 전체 162개의 키워드 중에서 설명변수(X)로 사용될 키워드와 반응변수(Y)로 사용될 키워드를 보여 준다.

Table 3. Keywords for variables

Variable	Keyword
Y	earthquake, fire
X	air, energy, gas, light, oil, speed, temperature, velocity, voltage, water, wind

본 논문에서는 대표적인 재난인 지진(earthquake)과 화재(fire)를 Y로 하고 지진과 화재에 영향을 미치는 키워드를 X로 선택하였다. 그림 1는 전체 X와 Y를 이용한 2-모드 네트워크의 시각화 결과이다.

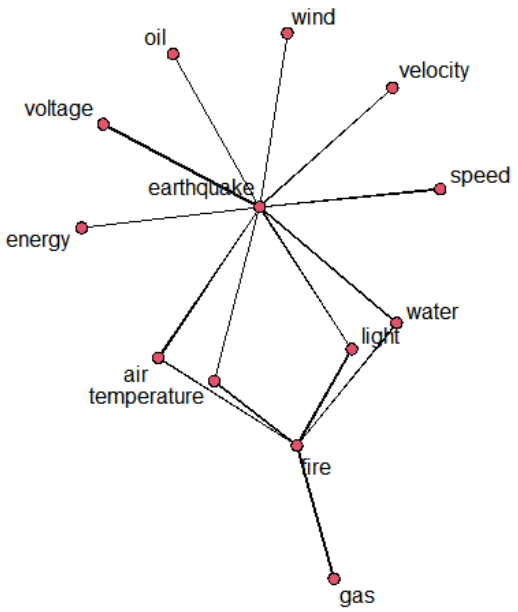

Fig. 1. Visualization of 2-mode network

식 (2)의 입력행렬을 통한 시각화를 통하여 반응변수로 사용된 earthquake와 fire를 중심으로 설명변수로 사용된

나머지 키워드의 연결 관계를 확인할 수 있다. 설명변수로 사용된 키워드 중에서 energy, voltage, oil, wind, velocity 그리고 speed는 earthquake와 직접 연결되어 있는 것을 알 수 있다. 반응변수로 사용된 키워드인 fire와 직접 연결된 키워드는 gas이다. 또한 air, temperature, light, water는 earthquake와 fire에 모두 연결되어 있다. 따라서 그림 1의 시각화 결과를 이용하여 표 4와 같이 반응변수에 따른 2개의 회귀 방정식을 구축할 수 있다.

Table 4. Regression equations by 2-mode network

Equation	Y	X
Eq. 1	earthquake	air, energy, light, oil, speed, temperature, velocity, voltage, water, wind
Eq. 2	fire	air, gas, light, temperature, water

반응변수로 사용된 earthquake와 fire에 직접 연결된 키워드는 모두 각 모형의 설명변수로 사용하였다. 각 모형에서 기존의 가우시안분포 기반의 GLM과 감마분포 기반의 GLM의 성능평가를 위하여 주어진 데이터에 평균이 0이고 분산이 σ인 정규분포로부터 생성된 난수의 절대값을 노이즈로 추가하였다. 왜냐하면 감마분포는 0보다 큰 연속형 값을 갖기 때문이다. 먼저 σ=1 인 경우의 성능평가 결과는 표 5와 6에서 보인다.

Table 5. Performance result (Y=earthquake, σ=1)

Model	AIC	BIC	log-L
Gaussian	38810	38903	-19393
Gaussian-noise	46980	47073	-23478
Gamma-noise	28921	29013	-14448

표 5의 결과는 Y가 earthquake이고 σ=1 인 경우이다. 모델이 Gaussian인 경우는 노이즈가 포함되지 않은 원래 데이터를 이용한 가우시안 기반 GLM을 이용한 결과이다. Gaussian-noise와 Gamma-noise는 원래 데이터에 평균이 0이고 σ=1인 정규분포로부터 생성된 난수의 절대값이 추가된 데이터를 이용하여 각각 가우시안 GLM과 감마 GLM을 수행한 결과를 나타낸다. Gaussian이나 Gaussian-noise에 비하여 Gamma-noise의 AIC와 BIC 값이 더 작게 나타남을 확인하였다. 또한 Gamma-noise의 로그우도가 가장 큰 것도 알 수 있다. 3가지 평가 측도 모두에서 Gamma-noise의 결과가 더 우수하게 나타남을 알 수 있다. 표 6는 Y가 fire이고 σ=1 인 경우이다.

Table 6. Performance result (Y=fire, σ=1)

Model	AIC	BIC	log-L
Gaussian	58175	58229	-29080
Gaussian-noise	61264	61318	-30625
Gamma-noise	30559	30613	-15272

표 5의 earthquake와 마찬가지로 반응변수가 fire인 경우에도 Gamma-noise의 성능이 다른 비교 모형들에 비하여 더 우수함을 알 수 있다. 다음으로 원래 데이터에 추가되는 노이즈의 표준편차를 σ=1.5로 증가시킨 경우의 모형간의 성능평가 결과가 표 7과 8에 나타나 있다.

Table 7. Performance result (Y=earthquake, σ=1.5)

Variable	AIC	BIC	log-L
Gaussian	38810	38903	-19393
Gaussian-noise	53639	53732	-26807
Gamma-noise	41504	41596	-20740

Y가 earthquake이고 σ=1.5인 표 7의 결과에서 AIC, BIC, 로그우도의 모든 평가측도에서 Gamma-noise의 성능은 Gaussian-noise에 비해서는 우수하지만 Gaussian에 비교해서는 약간의 성능저하를 확인할 수 있다. 표 8은 Y가 fire이고 노이즈의 표준편차 σ가 1.5인 경우이다.

Table 8. Performance result (Y=fire, σ=1.5)

Variable	AIC	BIC	log-L
Gaussian	58175	58229	-29080
Gaussian-noise	64443	64497	-32214
Gamma-noise	42684	42738	-21335

표 7의 결과와는 다르게 표 8에서는 σ=1인 표 5와 6의 결과와 마찬가지로 Gamma-noise의 성능이 가장 우수함을 알 수 있다. 따라서 원래 데이터에 추가되는 노이즈의 표준편차를 크게 한 경우에는 모형에 따라 약간의 성능차이가 나타났다. 다음의 표 9와 10에서는 노이즈의 표준편차를 σ=0.5로 작게 했을 때 모형의 성능평가 결과를 보여준다.

Table 9. Performance result (Y=earthquake, σ=0.5)

Variable	AIC	BIC	log-L
Gaussian	38810	38903	-19393
Gaussian-noise	41217	41310	-20596
Gamma-noise	8432	8524	-4203

Y가 earthquake이고 σ=0.5인 표 9의 결과에서 Gamma-noise의 AIC, BIC, 로그우도가 모두 다른 비교 모형들에 비해 우수하게 나타남을 확인할 수 있다. 특히 σ=1인 표 5의 결과에 비해서도 모형의 성능이 더 향상되었음을 알 수 있었다. 마지막으로 표 10은 Y가 fire이고 σ=0.5인 경우 모형의 성능평가 결과이다.

Table 10. Performance result (Y=fire, σ=0.5)

Variable	AIC	BIC	log-L
Gaussian	58175	58229	-29080
Gaussian-noise	59013	59067	-29499
Gamma-noise	10976	11030	-5481

표 9의 결과와 마찬가지로 표 10의 결과에서도 Gamma-noise의 성능이 가장 우수하게 나타남을 확인할 수 있었다. 본 논문의 전체 실험 결과를 통하여 우리는 특허 키워드 분석을 위하여 제안하는 방법의 모형성능이 다른 비교 모형들에 비하여 더 우수하게 나타남을 확인하였다. 특히 원래 데이터에 추가되는 노이즈의 표준편차는 $\sigma \leq 1$로 지정할 때 모형의 성능이 더 향상됨을 알 수 있었다.

V. Conclusions

본 논문에서 우리는 기존의 가우시안 확률분포 기반의 특허 키워드 분석 방법에 대한 성능향상을 위하여 2-모드 네트워크 시각화와 감마 확률분포 기반 GLM을 결합한 특허 키워드 분석 방법을 제안하였다. 목표기술을 결정하고 전세계 특허 데이터베이스로부터 관련된 특허문서를 수집하였다. 텍스트 마이닝의 전처리 과정을 수행하여 특허-키워드 행렬 형태의 정형화된 데이터를 구축하였다. 이 행렬로부터 키워드 간의 연관성을 기반으로 2-모드 네트워크 시각화가 구축되고 이로부터 감마 GLM을 위한 회귀식이 결정되어 최종 분석이 수행되었다.

실험 결과를 통하여 기존의 가우시안 분포 기반의 GLM에 비하여 제안 방법에 의한 모형의 성능향상을 확인할 수 있었다. 0부터 양의 무한대까지의 값을 가질 수 있는 특허 빈도 데이터의 특성을 고려했을 때 기존의 가우시안 분포보다 감마 분포에 의한 결과가 더 우수한 것을 확인하였다. 따라서 가우시안 확률분포에 기반한 GLM에 의존하는 기존의 특허 분석 방법을 다양한 확률분포 기반의 GLM으로 확장해야 할 필요성을 확인하였다.

본 논문의 연구결과는 기술분석이 필요한 연구개발기획, 기술사업화, 기술예측 등 다양한 기술경영 분야에서의 활용이 기대된다. 향후에는 키워드 뿐만 아니라 인용, 출원일, 기술코드 등 다양한 특허 데이터의 특성을 정교하게 반영할 수 있는 다양한 확률분포 기반의 모형화에 대한 지속적인 연구가 진행될 수 있을 것이다. 특히 본 연구에서는 재난 대처 기술에 한정하여 특허문서를 수집하고 분석하였다. 그러나 재난 대처 기술은 인공지능, 빅데이터, 사물인터넷 등 다양한 기술들과의 연관성을 고려해야 하기 때문에 향후 연구에서는 서로 연관성이 있는 기술들을 한꺼번에 고려하여 통합된 특허문서 데이터를 수집하고 분석하는 연구가 이루어질 것이다.

REFERENCES

[1] A. T. Roper, S. W. Cunningham, A. L. Porter, T. W. Mason, F. A. Rossini, and J. Banks, “Forecasting and Management of Technology” Hoboken, NJ, John Wiley & Sons, 2011.

[2] S. Jun, S. J. Lee, J. B. Ryu, and S. Park, “A novel method of IP R&D using patent analysis and expert survey,” Queen Mary Journal of Intellectual Property, Vol. 5, No. 4, pp. 474-494, October 2015.

[3] J. Kim, N. Kim, Y. Jung, and S. Jun, “Patent data analysis using functional count data model,” Soft Computing, Vol. 23, Iss. 18, pp. 8815–8826, September 2019.

[4] K. S. Alkaabi, and J. Yu, “A Study on the Integration Between Smart Mobility Technology and Information Communication Technology (ICT) Using Patent Analysis,” Journal of The Korea Society of Computer and Information, Vol. 24, No. 6, pp. 89-97, 2019.

[5] J. Park, and D. Kang, “Trend Analysis of Unmanned Technology Using Patent Information,” Journal of The Korea Society of Computer and Information, Vol. 22, No. 3, pp. 89-96, 2017.

[6] J. Park, “Trend Analysis of Artificial Intelligence Technology Using Patent Information,” Journal of The Korea Society of Computer and Information, Vol. 23, No. 4, pp. 9-16, 2018.

[7] S. Park, and S. Jun, “Technological Cognitive Diagnosis Model for Patent Keyword Analysis,” ICT Express, Vol. 6, pp. 57-61, 2020. https://doi.org/10.1016/j.icte.2019.09.004

[8] D. Uhm, J. Ryu, and S. Jun, “Patent Data Analysis of Artificial Intelligence Using Bayesian Interval Estimation,” Applied Sciences, Vol. 10, pp. 570, 2020. https://doi.org/10.3390/app10020570

[9] J. Choi, D. Jang, S. Jun, and S. Park, “A Predictive Model of Technology Transfer using Patent Analysis,” Sustainability, Vol. 7, No. 12, pp. 16175-16195, 2015.

[10] J. Choi, and S. Jun, “Big Data Smoothing and Outlier Removal for Patent Big Data Analysis,” Journal of The Korea Society of Computer and Information, Vol. 21, No. 8, pp. 77-84, 2016.
[11] J. Kim, J. Ryu, S. Lee, and S. Jun, “Penalized Regression Models for Patent Keyword Analysis,” Model Assisted Statistics and Applications-International Journal, Vol. 12. pp. 239-244, 2017.
[12] D. Hunt, L. Nguyen, and M. Rodgers, “Patent Searching Tools & Techniques” Hoboken, NJ, Wiley, 2007.
[13] I. Feinerer, K. Hornik, and D. Meyer, “Text mining infrastructure in R,” Journal of Statistical Software, Vol. 25, No. 5, pp. 1-54, March 2008.
[14] I. Feinerer, and K. Hornik, Package ‘tm’ Ver. 0.7-8 Text Mining Package, CRAN of R project, 2020
[15] S. Jun “Bayesian Count Data Modeling for Finding Technological Sustainability,” Sustainability, Vol. 10, No. 9, pp. 3220, 2018.
[16] K. P. Murphy, “Machine Learning: a probabilistic perspective” Cambridge MA, MIT Press, 2012.
[17] T. Hastie, R. Tibshirani, and J, Friedman, “The Elements of Statistical Learning, Data Mining, Inference, and Prediction” New York, Springer, 2017.
[18] D. C. Montgomery, E. A. Peck, and G. G. Vining, “Introduction to Linear Regression Analysis” Hoboken, New Jersey, John Wiley & Sons, 2012.
[19] C. T. Butts, “network: A Package for Managing Relational Data in R,” Journal of Statistical Software, Vol. 24, No. 2, pp. 1–36, 2008.
[20] C. T. Butts, D. Hunter, M. Handcock, S. Bender-deMoll, J. Horner, L. Wang, P. N. Krivitsky, B. Knapp, M. Bojanowski, and C. Klumb, Package ‘network’ Ver. 1.17.2, Classes for Relational Data, CRAN of R project, 2022.
[21] R. V. Hogg, J. W. Mckean, and A. T. Craig, “Introduction to Mathematical Statistics, eighth edition” Pearson, Essex, UK, 2020.
[22] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, “Bayesian Data Analysis, Third Edition” Boca Raton, FL, Chapman & Hall/CRC Press, 2013.
[23] R. V. Hogg, E. A. Tanis, and D. L. Zimmerman, “Probability and Statistical Inference ninth edition” Essex, England, Pearson, 2015.
[24] USPTO, The United States Patent and Trademark Office, http://www.uspto.gov, 2022.
[25] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, 2022.

Authors

Sunghae Jun received B.S., M.S., and PhD degrees from Department of Statistics, Inha University, Incheon, Korea in 1993, 1996, and 2001, respectively. He also received PhD degree from Department of Computer Science, Sogang University, Seoul, Korea in 2007. He is Professor in the Department of Big Data and Statistics, Cheongju University, Chungbuk, Korea. Also, He was visiting scholar in Department of Statistics, Oklahoma State University, Stillwater, Oklahoma, USA from 2009 to 2010. His current research interests include AI and data science.