electronics

MDPI

Article

Patent Keyword Analysis Using Regression Modeling Based on Quantile Cumulative Distribution Function

Sangsung Park and Sunghae Jun *

Department of Data Science, Cheongju University, Chungbuk 28503, Republic of Korea; hanyul@cju.ac.kr
* Correspondence: shjun@cju.ac.kr or statcs@gmail.com; Tel.: +82-10-7745-5677; Fax: +82-43-229-8432

Abstract: Patents contain detailed information of researched and developed technologies. We analyzed patent documents to understand the technology in a given domain. For the patent data analysis, we extracted the keywords from the patent documents using text mining techniques. Next, we built a patent document–keyword matrix using the patent keywords and analyzed the matrix data using statistical methods. Each element of the matrix represents the frequency of a keyword that occurs in a patent document. In general, most of the elements were zero because the keyword becomes a column of the matrix even if it occurs in only one document. Due to this zero-inflated problem, we experienced difficulty in analyzing patent keywords using existing statistical methods such as linear regression analysis. The purpose of this paper is to build a statistical model to solve the zero-inflated problem. In this paper, we propose a regression model based on quantile cumulative distribution function to solve this problem that occurs in patent keyword analysis. We perform experiments to show the performance of our proposed method using patent documents related to blockchain technology. We compare regression modeling based on a quantile cumulative distribution function with convenient models such as linear regression modeling. We expect that this paper will contribute to overcoming the zero-inflated problem in patent keyword analysis performed in various technology fields.

Keywords: patent keyword analysis; quantile cumulative distribution function; regression; patent document; patent–keyword matrix

Citation: Park, S.; Jun, S. Patent Keyword Analysis Using Regression Modeling Based on Quantile Cumulative Distribution Function. *Electronics* **2024**, *13*, 4247. https://doi.org/10.3390/electronics13214247

Academic Editors: Wentao Li, Huiyan Zhang, Tao Zhan and Chao Zhang

Received: 28 September 2024
Revised: 28 October 2024
Accepted: 28 October 2024
Published: 30 October 2024

1. Introduction

Patent keyword analysis (PKA) is important to technology management because a patent contains extensive and detailed information about the developed technology. Using the PKA results, we can build research and development (R&D) plans and strategies for the target technology. In general, for PKA, we extract technology keywords from patent documents using text mining techniques [1,2]. Using the extracted keywords, we construct a patent–keyword matrix for PKA based on statistics and machine learning algorithms. The matrix consists of elements representing the frequency values of keywords that occur in patents [1,3–5]. In most cases, this matrix has a sparse data structure that suffers from the zero-inflated problem [3–5]. This is because a keyword that is included in even just one patent document becomes one column in the matrix [3–5]. The sparse zero-inflated problem reduces the performance of PKA models [4,6]. As such, we have to solve the zero-inflated problem for PKA. Many existing studies rely on statistical models such as the zero-inflated Poisson and negative binomial models to solve the problem [5,7–11]. Recently, studies based on machine learning methods such as generative models have been conducted to solve the zero-inflated problem [4,5]. However, existing models have limitations in that model performance deteriorates as the proportion of zeros included in the data increases [3–5,7,8]. To solve this problem, we consider a regression model based on a quantile Cumulative Distribution Function (CDF) [12–14]. We call this model CDF-based Quantile Regression Model (QRM) in this paper. To verify the performance

of the CDF-based QRM, we perform experiments using patent documents related to blockchain technology.

The motivation for this research is to appropriately deal with the zero-inflated problem that occurs in patent keyword data analysis. In particular, we study a method to overcome the extreme zero-inflated problem, where the proportion of zeros in the given data exceeds half. Since the extreme zero-inflated problem is difficult to solve even with existing statistical zero-inflated models, we need to find new methods to solve it.

The remainder of this paper is organized as follows. We survey works related to our research such as regression and zero-inflated models in Section 2. In Section 3, we present the theoretical explanation of our proposed method and the analysis process step by step. In addition, we present the performance evaluation indexes of comparative models in this section. Next, we show the improved performance and validity of our proposed method from the experimental results using patent documents related to blockchain technology in Section 4. In this section, we compare the model performance of the CDF-based QRM with traditional linear regression and statistical zero-inflated models. In the Section 5, we illustrate how the proposed method can be applied to practical tasks in various domains. Lastly, we provide the conclusions and future works related to our research in Section 6.

2. Related Works

Patent analysis has been performed in various technology domains such as photovoltaic, medicine, mountain logistics, climate change, artificial intelligence (AI), surgery, and energy [15–21]. This is because when developers register a technology they have developed as a patent, they are guaranteed exclusive rights to use their technology for a certain period of time. Therefore, we analyze patents to understand these technologies. Also, we use the results of patent analysis for technology management such as Research and Development strategy development. PKA, which we propose in this paper, is also a field of patent analysis. PKA mainly extracts technology keywords from the abstracts and claims contained in collected patent documents and analyzes them. In this process, we use text mining and various data analysis methods based on statistics and machine learning.

The regression model is very popular in machine learning as well as in statistics [12,22–24]. This model consists of independent and dependent variables called X and Y, respectively [22]. Regression analysis is statistical modeling that explores relationships between variables [24]. We can predict Y for a given X using regression analysis [19]. Figure 1 shows a process of regression modeling [23].

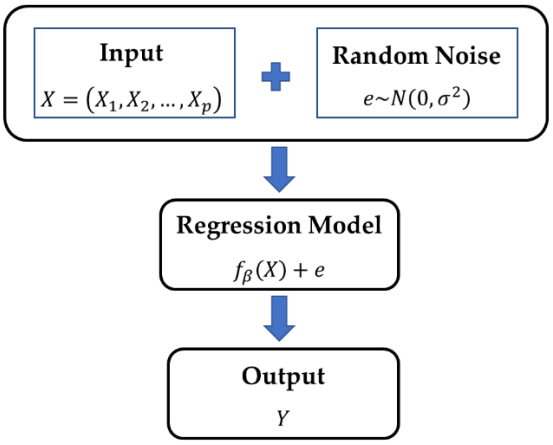

Figure 1. Regression modeling process.

We assume that the response variable Y adds the error e to the explanatory variables X Using X and Y, we create the linear regression model (LRM) as follows.

$$Y = f_\beta(X) + e \tag{1}$$

In Equation (1), $f_\beta(X)$ is $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$, and we estimate the model parameters, $\beta = (\beta_0, \beta_1, \beta_2, \ldots \beta_p)$ that minimize the error using the least squared loss function [12,24]. The error represents a random noise included in observed data and follows a normal distribution with a mean (μ) = 0 and variance = σ^2. This model has provided good performance in exploring the relationship between X and Y in most data, including errors with a mean of 0 and equal variance [24]. However, we have difficulty in using the LRM when the given data does not satisfy the model assumptions [9–11]. In particular, if the given data has many extreme values, we cannot use the LRM [4–6]. To solve the problems of LRM, we can consider the QRM [13,25]. Quantile regression aims to model the impact of explanatory variables on the quantile of the response variable. The QRM finds the conditional quantile of Y just as the regression based on the least square method estimates the conditional mean of Y [9]. We can apply both continuous and count data to QRM. QRM is a model that can be used when the given data do not satisfy the normality assumption and are asymmetric or contain many outliers. In the PKA, we found that the patent–keyword matrix contains zero-inflated data that is sparse and asymmetric. As such, we propose a method to analyze the patent keywords using QRM. In addition, we consider the CDF for our PKA model based on QRM because we aim to predict the specific quantile of each patent keyword.

In statistics, the zero-inflated model is typically used to analyze data that contain a lot of zeros [8,10,11]. This model has been used to solve the zero-inflated problem that occurs in various domains [26–29]. The zero-inflated model based on statistics is defined as follows [9]:

$$P(X = x) = \begin{cases} \pi + (1 - \pi) f(x), & x = 0 \\ (1 - \pi) f(x), & x > 0 \end{cases} \quad (2)$$

In Equation (2), $f(x)$ is a density of random variable $X = x$. In the statistical zero-inflated model, the probability model of X is separated into two parts of zero and non-zero [9–11]. The π represents the probability of zero occurrence. Although the statistical zero-inflated models have been used to overcome the problem that arises in various data analysis processes, they have shown a problem in that model performance deteriorates as the proportion of zeros in the data increases [3–5]. Therefore, in this paper, we propose a PKA method using QRM for analyzing patent keyword data with a high zero ratio.

3. Proposed Method

3.1. Patent–Keyword Matrix

The reason we analyze patent keywords is because patent keywords represent technologies. Many governments and companies around the world are working hard to establish R&D strategies for new and promising technology development in order to survive in the fierce technological competition. For this purpose, understanding of technology is essential, and one of the effective methods for understanding technology is patent keyword analysis. In order to analyze patent keywords, we construct a patent–keyword matrix from patent documents. However, since a significant portion of the elements in this matrix are zero, existing data analysis methods have limitations. For example, if the given data contains too many zeros, the explanatory and predictive power of models built using this data will be reduced. To solve this problem, in this paper, we propose an analytical method for patent keyword data. The patent keyword data are generated from the title and abstract parts of patent documents. In general, we search the patent documents related to the target technology from various patent databases in the world using a keyword search equation. For our PKA, we preprocess the searched patent documents using text mining techniques as follows [1,2].

(TM.1) Searching patent documents related to target technology.
(1-1) Using keyword searching equation, we collect the patent documents related to the target technology.

(1-2) By examining all the retrieved patents, we select valid patents that can be used for analysis.
(TM.2) Building structured patent data by text mining.
(2-1) Using tokenization and normalization, we preprocess the patent documents to create the corpus.
(2-2) By extracting keywords from the corpus, we construct a patent–keyword matrix.

In the first text mining (TM.1) step, we determine the target technology for PKA. In this step, we collect the patent documents related to the target technology from various patent databases. In addition, we select the valid patents from the collected patent documents. Using tokenization and normalization methods such as stemming, lemmatization, lowercasing, and removing stopwords, we preprocess the valid patent documents and make a corpus of them in the second text mining (TM.2) step. Finally, we extract the patent keywords from the corpus and construct the patent–keyword matrix. The rows and columns are patent documents and keywords from the vocabulary, respectively. Also, the matrix values are the frequency values of each keyword in a document. Therefore, through the text mining process of making text corpus, parsing, and constructing text database, we build a patent–keyword matrix [1,3,4]. We define this matrix M using Equation (3).

$$M = \left(Freq_{ij}\right),\ i = 1,\ 2, \ldots, p\ j = 1,\ 2, \ldots,\ n \tag{3}$$

where p and n are the numbers of keywords and patent documents, respectively. Also, $Freq_{ij}$ is frequency value of the jth keyword occurring in the ith patent. The observed data of $Freq_{ij}$ is distributed to Poisson probability distribution with parameter λ as follows.

$$X = Freq_{ij},\ X\ \sim\ Poisson(\lambda),\ \lambda > 0,\ x = 0,\ 1,\ 2, \ldots \tag{4}$$

In Equation (4), X has values greater than or equal to zero, but most X values are zeros. This is one cause of the deteriorating performance of statistical analysis models in PKA. This is the problem we aim to solve in this paper. Next, we normalize the frequency values of the patent–keyword matrix as follows.

$$Freq_{ij\_nor} = \frac{Freq_{ij} - Min}{Max - Min} \tag{5}$$

Max and Min represent the maximum and minimum among all values of $Freq_{ij}$ in Equation (5). According to Equation (5), the range of values that $Freq_{ij\_nor}$ can have changes the values of $Freq_{ij}$ from an integer greater than 0 to a real number between 0 and 1. In most cases, the patent–keyword matrix has the zero-inflated problem. That is, most of the frequency values are zeros. This becomes a factor that seriously reduces the performance of the analysis model. Therefore, we have to deal with this problem for PKA. We try to solve this problem using the CDF-based QRM in our study.

3.2. Quantile Regression Modeling Based on Cumulative Distribution Function for PKA

The patent–keyword matrix is asymmetric and sparse because of the zero-inflated problem of matrix elements. Most of the elements in the matrix have the value zero. Therefore, the matrix has a very imbalanced data structure. Because of the characteristics of the patent–keyword matrix with such an asymmetric structure, we have difficulty in analyzing patent keywords using statistical techniques. This problem reduces the analytical performance of statistical methods and machine learning algorithms. To solve the problem, we propose a method of PKA using CDF-based QRM. The CDF is defined as follows [9,12].

$$F(y) = P(Y \leq y) \tag{6}$$

In Equation (6), $F(y)$ is the value of CDF of $Y = y$ and is computed by the probability of $Y \leq y$. Also, the q-th quantile of Y is a value of between 0 and 1. Therefore, we have to

change the value of Y to a (0,1) interval. In our study, we normalized the frequency value of response keywords to a real value between 0 and 1. This approach is similar to PKA using beta regression modeling. The probability distribution corresponding to a random variable Y with support between 0 and 1 is the beta distribution [9,24]. Though the regression analysis model based on the beta distribution can be used for PKA, as the zero-inflated problem becomes more severe, the beta regression model also shows limitations in model performance like existing regression models [4–6]. Therefore, we use CDF-based QRM to analyze the patent–keyword matrix data. This models the CDF of the response variable (keyword) Y and makes predictions for specific quantiles of Y. In other words, we can use the model to estimate the probability that the response variable is below a certain value. Because the CDF-based QRM has robust characteristics that are not significantly affected by extreme values such as outliers, we use this model for analyzing patent keywords with the sparsity problem of zero inflation. The CDF with a location parameter μ and a dispersion parameter σ is defined with the following equation [13,25].

$$\mathrm{G}(X = x, \mu, \sigma) = F\Big[U\Big(H^{-1}(x), \mu, \sigma\Big)\Big],\ 0 \le x \le 1,\ -\infty < \mu < \infty,\ 0 < \sigma \tag{7}$$

In Equation (7), the random variable X has the support (0,1) and two parameters, μ and σ. F and H are a CDF and an invertible CDF with supports D_1 and D_2, respectively. U is a suitable transformation from $D_1(-\infty, \infty)$ to $D_2(0, \infty)$ for applying μ and σ. In addition, $H^{-1}(x)$ is a corresponding quantile function from D_1 to D_2. Using this CDF quantile family in Equation (7), we carry out the CDF-based QRM for our PKA. To estimate the parameters, we carry out maximum likelihood estimation (MLE) for all parameters based on a gradient [12,13,23,25]. For the CDF in Equation (7), the probability density function (PDF) is defined as in Equation (8) [24,25].

$$G(X = x, \mu, \sigma) = \frac{q(x) f\left(\frac{H^{-1}(x) - \mu}{\sigma}\right)}{\sigma},\ 0 \le x \le 1,\ -\infty < \mu < \infty,\ 0 < \sigma \tag{8}$$

where $f(x)$ and $q(x)$ are the PDF corresponding to F and the quantile density function corresponding to H^{-1}. We differentiate the log of G with respect to μ and σ and dropping q. As such, the regression model has two sub models as follows [13]:

$$L_\mu(\hat{\mu}) = x^T \beta \tag{9}$$

$$L_\sigma(\hat{\sigma}) = z^T \delta \tag{10}$$

where Equations (9) and (10) are the models for location (μ) and dispersion (σ), respectively. x and z are the vectors of predictors. Also, β and δ are the vectors of coefficients. In this paper, we use the link functions of identity and log for L_μ and L_σ. In the next section, we conduct experiments using patents related to blockchain technology to evaluate the model performance. We illustrate our PKA process of CDF-based QRM as follows.

(Step.1) Patent data collection and preprocessing.
(1-1) Searching patent documents from patent databases.
(1-2) Extracting keywords from searched patents using text mining.
(1-3) Constructing the patent–keyword matrix from extracted keywords.
(Step.2) Data preparation and normalization.
(2-1) Determining response and explanatory variables (keywords).
(2-2) Normalizing response variable to a (0,1) interval.
(Step.3) Data analysis.
(3-1) Building the CDF-based QRM.
(3-2) Evaluating model performance using loglikelihood, AIC, and BIC.

Our proposed PKA method consists of three steps. In Step.1, we collect patent documents related to the target technology from patent databases such as the United States Patent and Trademark Office (USPTO). We extract keywords from the collected patent

documents using various text mining techniques. For patent analysis, we construct a matrix consisting of patents and keywords for rows and columns, respectively. Also, each element of the matrix is the frequency of occurrence of each keyword in the patent document. Next, we select response and explanatory variables according to the target technology and the aim of the patent analysis. For CDF-based QRM, we normalize the value of the response variable to a (0,1) interval. In the final step, we use the CDF-based QRM to analyze the patent–keyword matrix, which is sparse and has the zero-inflated problem. In this paper, we compared the CDF-based QRM with the LRM. This is because the LRM is widely used in the field of keyword data analysis. We used three indexes to evaluate the performance between the comparative models. First, we use the loglikelihood, defined as in Equation (11) [9,13,24].

$$L(\theta|x) = \sum_{i=1}^{n} log(f(x_i|\theta)) \tag{11}$$

where θ and n are the model parameter and data size. $f(x_i|\theta)$ is a joint probability density (or mass) function of x_i given θ. The larger the loglikelihood value of a model is, the better the model fits the data. Next, we consider the Akaike information criterion (AIC) to evaluate the performance of model fitting. This is represented in the following equation [23].

$$AIC = -2L(x|\hat{\theta}) + 2k \tag{12}$$

In Equation (12), $\hat{\theta}$ is the maximum likelihood estimate of θ and $L(x|\hat{\theta})$ is the maximum loglikelihood function given x. Also, k is the number of explanatory variables. The smaller the AIC value, the better the fitting performance of the model. Lastly, we apply the Bayesian information criterion (BIC) index to evaluate the performance of comparative models. The value of BIC is computed as follows [23]:

$$BIC = -2L(x|\hat{\theta}) + k \times log(n) \tag{13}$$

The BIC is an index that adds consideration to data size n to the AIC in Equation (13). As with the AIC, in the case of the BIC, the smaller this value is, the better the model performance is. In this paper, we compared the proposed QRM with LRM and a zero-inflated model in terms of explanatory and predictive power. Loglikelihood is an index that measures the explanatory power of the model, and the AIC and BIC are indexes that compare the predictive power between models.

4. Experiments and Results

The experiments were carried out using practical patent documents to illustrate how the proposed method can be applied to real fields. We collected patent documents related to blockchain technology from world patent databases [30,31]. Blockchain technology has been developed by relying on the blockchain-related technologies such as bitcoin and cryptocurrency. So, in this experiment, we provide the technological relations between blockchain technology and other related technologies based on the keywords of blockchain, access, authentication, bitcoin, cryptocurrency, databank, distributor, encash, ledger, network, and secretkey. In this paper, we determined blockchain technology as our target domain. Blockchain is defined as a technology for securely managing data across distributed systems [6]. We select the keyword blockchain as the response variable and use the remaining ten extracted keywords (access, authentication, bitcoin, cryptocurrency, databank, distributor, encash, ledger, network, and secretkey) for explanatory variables ($X_1, X_2, \ldots, X_{10}$). Figure 2 shows the process of our proposed modeling of PKA.

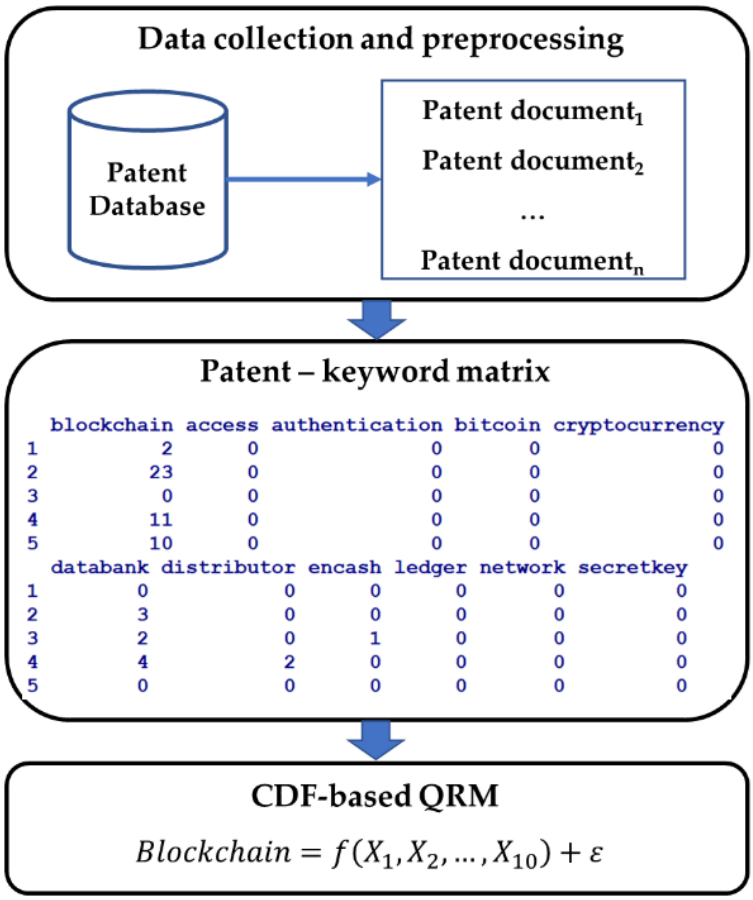

Figure 2. Patent keyword analysis process.

First, we collected the patent documents related to blockchain technology using keyword search expression from patent databases across the world [30,31]. Next, we chose the valid patents representing blockchain technology and preprocessed the valid patent documents. In our experiments, we used the R project as a tool for statistical analysis. R is a free, open-source piece of software that supports statistical analysis and visualization [32]. The current version of R has been upgraded to 4.4.2. Up until now, R has been widely used for statistical analysis of data generated in various fields [33]. We also used the tm package of R for text mining [1]. This package provides many functions for preprocessing of text data using natural language processing [1,2]. Lastly, we used the cdfquantreg package of R for QRM [14]. In addition, using the functions provided in the R base module and the pscl package, we carried out performance evaluation between the proposed model and the comparative models [32,34]. The elements of this matrix are the frequency values of the keywords occurring in the patent documents. This is structured data that can be used in CDF-based QRM. Also, we determined the keyword of blockchain for the dependent variable and used the other keywords for independent variables in this experiment. To select the patent keywords for blockchain technology, we considered the results of keyword extraction from previous research related to blockchain technology analysis [32]. Therefore, we determined one response variable (blockchain) and ten explanatory variables (access, authentication, bitcoin, cryptocurrency, databank, distributor, encash, ledger, network, and secretkey). We used the R project and package for our experiment [1,2,14,32–34]. Table 1 shows the summary statistics of the patent keywords.

Table 1. Summary statistics of blockchain patent keywords.

Keyword	Min	Median	Mean	Max
blockchain	0	2	4.4140	38
access	0	0	0.2606	11
authentication	0	0	0.4763	16
bitcoin	0	0	0.1701	20
cryptocurrency	0	0	0.2115	11
databank	0	0	1.7800	24
distributor	0	0	0.5626	14
encash	0	0	0.1058	4
ledger	0	0	0.8892	26
network	0	0	0.9332	16
secretkey	0	0	0.5144	10

In the results in Table 1, we found that the patent–keyword matrix data is very sparse and zero-inflated because most elements of the matrix are zeros. The median values of most keywords were also zero. Therefore, we have difficulty analyzing the patent keyword data using traditional data analysis methods. To overcome the problem, we proposed patent keyword analysis using CDF-based QRM in this paper. In the CDF-based QRM, the response variable must have real numbers between 0 and 1. So, we changed the values of blockchain keyword by the following normalization.

$$Blockchain_{normalization} = \frac{Blockchain - Min(Blockchain)}{Max(Blockchain) - Min(Blockchain)} \quad (14)$$

Using the Equation (14), the values of the response variable are changed to numerical values in the interval (0,1). The model of patent keyword analysis consists of one response variable of the keyword blockchain and ten explanatory variables of all keywords except blockchain as follows.

Y: blockchain
$X_1, X_2, \ldots, X_{10}$: access, authentication, bitcoin, cryptocurrency, databank, distributor, encash, ledger, network, secretkey

Using the indexes of (11), (12), and (13), we compared the performance between CDF-based QRM and LRM. Table 2 shows the results of model performance between the compared models according to loglikelihood, AIC, and BIC. In this paper, we compared our proposed QRM with LRM and the zero-inflated model in terms of explanatory and predictive power. Loglikelihood is an index that measures the explanatory power of the model, and AIC and BIC are indexes that compare the predictive power between models.

Table 2. Results of performance evaluation between comparative models.

Model	Loglikelihood			AIC			BIC		
	QRM	LRM	ZIP	QRM	LRM	ZIP	QRM	LRM	ZIP
access	1370.29	549.57	−3621.28	−2734.59	−1093.14	7250.55	−2719.36	−1077.91	7270.85
authentication	1370.81	548.26	−3627.43	−2735.61	−1090.51	7262.86	−2720.39	−1075.29	7270.85
bitcoin	1397.82	553.43	−3610.05	−2789.63	−1100.85	7228.10	−2774.41	−1085.63	7248.40
cryptocurrency	1402.22	557.10	−3598.48	−2812.43	−1108.21	7204.97	−2797.20	−1092.98	7225.27
databank	1383.69	561.47	−3599.97	−2761.38	−1116.94	7207.93	−2746.16	−1101.71	7228.23
distributor	1370.23	548.67	−3625.33	−2734.45	−1091.35	7258.66	−2719.23	−1075.13	7278.96
encash	1370.55	549.89	−3623.16	−2735.11	−1093.77	7254.32	−2719.88	−1078.55	7274.62
ledger	1390.69	573.09	−3528.22	−2775.38	−1140.18	7064.44	−2760.16	−1124.95	7084.74
network	1374.97	555.71	−3612.41	−2743.93	−1105.43	7232.82	−2728.70	−1090.20	7253.12
secretkey	1382.74	551.14	−3622.10	−2759.47	−1096.27	7252.20	−2744.25	−1081.05	7272.50
All keywords	1486.48	610.57	−3429.08	−2948.96	−7797.14	6902.16	−2888.06	−1136.24	7013.81

In Table 2, to compare the performance of CDF-based QRM and LRM, we built simple models consisting of one keyword each and a full model using all keywords. First, the loglikelihood result shows that the loglikelihoods of CDF-based QRM for all keywords are larger than those of LRM. This shows that the results of patent keyword analysis using CDF-based QRM are better than those of the LRM. Next, in the comparison results based on AIC, the AIC values of CDF-based QRM are smaller than those of LRM for both the model using all keywords as well as the model using each keyword. We illustrate that CDF-based QRM is superior to LRM from the AIC perspective. Lastly, we compared the BIC values between CDF-based QRM and LRM. In Table 2, we can see that the BIC values of CDF-based QRM are larger than those of LRM. This means that the model performance of CDF-based QRM is better than LRM. Therefore, we show the validity of our proposed approach to patent keyword analysis from the comparison results by loglikelihood, AIC, and BIC.

The last column of each index that evaluates the performance of the model presents the results of the analysis using the statistical zero-inflated model. In this paper, we used the zero-inflated Poisson (ZIP) model as a statistical zero-inflated model [10,11]. This model uses the Poisson distribution as the probability function of the statistical zero-inflated model. The following shows the ZIP model [10,11].

$$\mathrm{P}(\mathrm{X}=\mathrm{x}) = \begin{cases} \pi + (1-\pi)e^{-\lambda}, \mathrm{x}=0 \\ (1-\pi)\dfrac{e^{-\lambda}\lambda^{x}}{\mathrm{X}!}, \mathrm{x}>0 \end{cases} \tag{15}$$

Equation (15) uses the probability function of the Poisson distribution as $f(x)$ in Equation (3). In Equation (15), the λ is the parameter of Poisson distribution. In all indexes of loglikelihood, AIC, and BIC, we confirmed that model performance of ZIP is inferior to that of QRM or LRM. This is because the proportion of zeros included in the patent–keyword matrix data exceeds half, as we confirmed in Table 1. Therefore, we could confirm that our QRM is superior to the LRM or ZIP models. Finally, we represent the estimated parameter and p-value of each keyword in Table 3.

Table 3. Estimated parameter and p-value of each keyword.

Keyword	Estimated Parameter	p-Value
access	0.4714	0.3091
authentication	−0.2966	0.3479
bitcoin	−3.0515	<0.0001
cryptocurrency	−3.7862	<0.0001
databank	0.4457	<0.0001
distributor	−0.5016	0.1500
encash	0.9106	0.3098
ledger	0.6810	<0.0001
network	0.7086	0.0007
secretkey	−2.6537	<0.0001

Depending on the keyword, we found that some keywords have a positive impact on blockchain while others have a negative impact. Additionally, through the result of p-value, we confirmed that the keywords bitcoin, cryptocurrency, databank, ledger, network, and secretkey have a statistically significant impact on blockchain technology because the p-values of these keywords are less than 0.05 at the 95% confidence level. We can apply the results in Table 3 to various technology management areas such as R&D planning. From the result of Table 3, we constructed a technology diagram of blockchain in Figure 3.

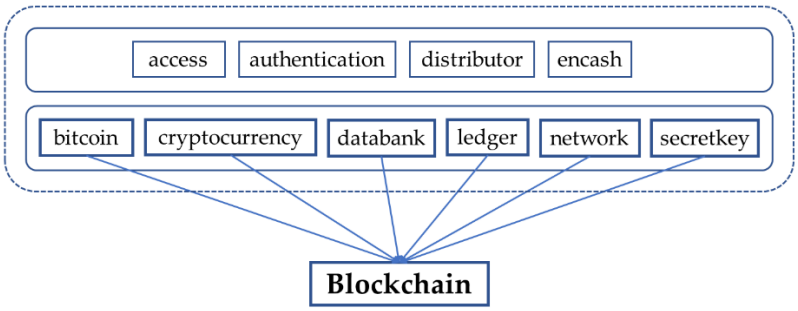

Figure 3. Blockchain technology diagram.

Among the 10 keywords related to blockchain, we can see that the keywords of bitcoin, cryptocurrency, databank, ledger, network, and secretkey have a statistically significant effect on blockchain. Therefore, we can see that technologies based on these keywords are primarily necessary for the development of blockchain technology. We expect that these results will contribute to R&D planning for blockchain technology development in countries and companies.

5. Discussion

From the result of Table 1, we found that the patent keyword data related to blockchain technology exhibit the zero-inflated problem. So, we used the proposed method to solve the problem. From the results in Table 2, we confirmed that the performance of QRM is better than those of the LRM and ZIP. Therefore, we estimated the model parameters and their p-values using the QRM in Table 3. Lastly, using the results in Table 3, we constructed a technology diagram of blockchain in Figure 3. From the results in Figure 3, we found that the sub-technologies based on the keywords of bitcoin, cryptocurrency, databank, ledger, network, and secretkey have a significant influence on the development of blockchain technology. In this paper, we applied the proposed method to analyze patent keyword data related to blockchain technology. From our experimental results, we showed how our method could be applied to real technology domains. Although the practical technology domain we used is blockchain, we believe that our proposed method can be extended to other technology fields. Once the target technology is determined, patent keyword analysis can be performed according to each step of the method proposed in this paper. Through this, we can conduct patent analysis necessary for R&D planning, new product development, technology forecasting, and technology innovation required in technology management.

In addition, we derived the QRM based on CDF to analyze the patent keywords. The patent–keyword matrix, which is usually used for patent keyword analysis, contains a large number of zeros, making it difficult for us to use existing linear models. If there are too many zero values, the zero values dominate the model building, which reduces the model performance. To solve this problem of zero inflation, we studied and proposed the CDF-based QRM in this paper.

In our study, we tried to identify the relationship structure between technologies through the PKA. Figure 3 was the final result obtained from our study. The technology diagram of Figure 3 provides a list of keywords that are statistically significant for blockchain technology. Therefore, in order to effectively develop blockchain technology, we must pay attention to detailed technologies based on these keywords. However, the results in Figure 3 do not provide any predictive information about the future technology of blockchain. In order to continuously develop blockchain technology, we need to predict the technology of blockchain. In addition, predicting the next technology related to the target technology will also be very meaningful in understanding the technology. To this end, it would also be meaningful to study how to use machine learning methods to predict the next behavior of animals [35]. Just as past patterns of animal behavior can be analyzed

to predict future behavior, past patterns of technological development can be modeled to predict future technologies.

6. Conclusions

This paper presents a statistical model in order to solve the zero-inflated problem. We collected patent documents related to blockchain technology and analyzed them using a statistical data analysis method. Blockchain technology is a data management technology with distributed secure applications in various domains such as the financial field of Bitcoin. This technology is based on decentralization, immutability, transparency, and security. In this process, we constructed a patent–keyword matrix using preprocessed data for statistical analysis. Each element of this matrix is a frequency value of a keyword's occurrence in a patent document. Because most of the elements in this matrix are zero, we had difficulty analyzing this matrix using statistical analysis methods including the zero-inflated model. Therefore, we proposed a method of PKA to overcome the zero-inflated problem in the preprocessed patent data. Compared to existing single models such as LRM, we considered an analysis model consisting of two sub models representing location and dispersion. In addition, we changed the value of the response variable to a (0,1) interval. This is the concept of the CDF-based QRM.

In our experiment, we compared the model performance of the CDF-based QRM with LRM and ZIP to show the improved performance of our model. We searched the patents related to blockchain technology. The analytical results provided by the CDF-based QRM, LRM, and ZIP were evaluated using loglikelihood, AIC, and BIC. We found that all experimental results of the CDF-based QRM were better than those of the LRM and ZIP. Therefore, we showed the validity of the CDF-based QRM for our PKA. In the CDF-based QRM, we normalized the scale of the response variable y to solve the zero-inflated problem and confirmed the improved performance of the proposed method.

In this paper, the proposed model was used to finally select technology keywords that have a statistically significant impact on blockchain technology. We had difficulty identifying technological relationships between patent keywords using our proposed model. However, understanding the interrelationship structure between the sub-technologies required for blockchain technology development is an important task in understanding this technology. This part represents the limitations of our study. To overcome the limitation of our proposed model, we considered social network analysis (SNA) and Bayesian learning. In our future works, we apply SNA to our CDF-based QRM to make a technology diagram representing the technological relations between the patent technology keywords. In addition, we will apply Bayesian learning to the CDF-based QRM. We call this Bayesian learning for QRM. In this model, we assume the prior distributions for the parameters of the QRM model. This learning model updates the model parameters using the given data. That is, we will be able to improve the QRM performance of explanatory and predictive power using the Bayesian learning process whenever new data are added. The prior distribution of the parameters is updated by combining it with the likelihood function of newly observed data to form the posterior distribution of the parameters. That is, the parameters become random variables with probability distribution functions rather than fixed values and can be effectively used in the analysis of a zero-inflated patent–keyword matrix. Our research is expected to contribute to various fields by improving understanding of technology and finding relationships between detailed technologies through our PKA.

Author Contributions: S.P. designed this research and collected the dataset for the experiment. S.J. analyzed the data to show the validity of this paper, wrote the paper, and performed all the research steps. In addition, all authors cooperated with each other in revising the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Feinerer, I.; Hornik, K. *Package 'tm' Version 0.7-12, Text Mining Package*; CRAN of R Project; R Foundation for Statistical Computing: Vienna, Austria, 2024.
2. Feinerer, I.; Hornik, K.; Meyer, D. Text mining infrastructure in R. *J. Stat. Softw.* **2008**, *25*, 1–54. [CrossRef]
3. Park, S.; Jun, S. Zero-Inflated Patent Data Analysis Using Compound Poisson Models. *Appl. Sci.* **2023**, *13*, 4505. [CrossRef]
4. Uhm, D.; Jun, S. Zero-Inflated Patent Data Analysis Using Generating Synthetic Samples. *Future Internet* **2022**, *14*, 211. [CrossRef]
5. Jun, S. Zero-Inflated Text Data Analysis using Generative Adversarial Networks and Statistical Modeling. *Computers* **2023**, *12*, 258. [CrossRef]
6. Park, S.; Jun, S. Sustainable Technology Analysis of Blockchain Using Generalized Additive Modeling. *Sustainability* **2020**, *12*, 10501. [CrossRef]
7. Wagh, Y.S.; Kamalja, K.K. Zero-inflated models and estimation in zero-inflated Poisson distribution. *Commun. Stat.-Simul. Comput.* **2018**, *47*, 2248–2265. [CrossRef]
8. Feng, C.X. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *J. Stat. Distrib. Appl.* **2021**, *8*, 8. [CrossRef]
9. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2013.
10. Hilbe, J.M. *Negative Binomial Regression*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2011.
11. Hilbe, J.M. *Modeling Count Data*; Cambridge University Press: Cambridge, UK, 2014.
12. Hogg, R.V.; Mckean, J.W.; Craig, A.T. *Introduction to Mathematical Statistics*, 8th ed.; Pearson: Essex, UK, 2020.
13. Shou, Y.; Smithson, M. cdfquantreg: An R Package for CDF-Quantile Regression. *J. Stat. Softw.* **2019**, *88*, 1–30. [CrossRef]
14. Shou, Y.; Smithson, M. *Package 'cdfquantreg' Version 1.3.1-2, Quantile Regression for Random Variables on the Unit Interval Package*; CRAN of R Project; R Foundation for Statistical Computing: Vienna, Austria, 2023.
15. Ding, J.; Du, D.; Duan, D.; Xia, Q.; Zhang, Q. A network analysis of global competition in photovoltaic technologies: Evidence from patent data. *Appl. Energy* **2024**, *375*, 124010. [CrossRef]
16. Shi, R.; Chai, K.; Wang, H.; Guo, S.; Zhai, Y.; Huang, J.; Yang, S.; Li, J.; Zhou, J.; Qiao, C.; et al. Comparative effectiveness of five Chinese patent medicines for non-alcoholic fatty liver disease: A systematic review and Bayesian network meta-analysis. *Phytomedicine* **2024**, *135*, 156124. [CrossRef]
17. Teshome, M.B.; Podrecca, M.; Orzes, G. Technological trends in mountain logistics: A patent analysis. *Res. Transp. Bus. Manag.* **2024**, *57*, 101202. [CrossRef]
18. Elsen, M.; Tietze, F. Contributions from low- and middle-income countries to the development of climate change adaptation technologies: A patent analysis. *Technol. Forecast. Soc. Change* **2024**, *209*, 123660. [CrossRef]
19. Zhao, X.; Wu, W.; Wu, D. How does AI perform in industry chain? A patent claims analysis approach. *Technol. Soc.* **2024**, *79*, 102720. [CrossRef]
20. Patel, M.S.; Franceschelli, D.; Grossbach, A.; Zhang, J.K.; Mercier, P.A.; Mattei, T.A. Top 50 Spine Surgery Publications Most Cited by Patents: A Bibliometric Analysis Focused on Research Driving Innovation. *World Neurosurg.* **2024**, *191*, 234–244. [CrossRef]
21. Ovsyannikov, I.R.; Zhdaneev, O.V. Forecast of innovative activity in key areas of energy transition technologies based on analysis of patent activity. *Int. J. Hydrogen Energy* **2024**, *87*, 1261–1276. [CrossRef]
22. Bruce, P.; Bruce, A.; Gedeck, P. *Practical Statistics for Data Scientists*, 2nd ed.; O'Reilly Media: Sebastopol, CA, USA, 2020.
23. Theodoridis, S. *Machine Learning a Bayesian and Optimization Perspective*; Elsevier: London, UK, 2015.
24. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
25. Smithson, M.; Shou, Y. CDF-quantile distributions for modelling random variables on the unit interval. *Br. J. Math. Stat. Psychol.* **2017**, *70*, 412–438. [CrossRef]
26. Chafamo, D.; Shanmugam, V.; Tokcan, N. C-ziptf: Stable tensor factorization for zero-infated multi-dimensional genomics data. *BMC Bioinform.* **2024**, *25*, 323. [CrossRef]
27. Yirdaw, B.E.; Debusho, L.K.; Samuel, A. Application of longitudinal multilevel zero infated Poisson regression in modeling of infectious diseases among infants in Ethiopia. *BMC Infect. Dis.* **2024**, *24*, 927. [CrossRef]
28. Zhou, W.; Huang, D.; Liang, Q.; Huang, T.; Wang, X.; Pei, H.; Chen, S.; Liu, L.; Wei, Y.; Qin, L.; et al. Early warning and predicting of COVID-19 using zero-infated negative binomial regression model and negative binomial regression model. *BMC Infect. Dis.* **2024**, *24*, 1006. [CrossRef]
29. Ren, J.; Loughnan, R.; Xu, B.; Thompson, W.K.; Fan, C.C. Estimating the total variance explained by whole-brain imaging for zero-inflated outcomes. *Commun. Biol.* **2024**, *7*, 836. [CrossRef]
30. KIPRIS. Korea Intellectual Property Rights Information Service. Available online: www.kipris.or.kr (accessed on 1 July 2023).
31. USPTO. The United States Patent and Trademark Office. Available online: http://www.uspto.gov (accessed on 1 June 2023).
32. R Development Core Team. R: A Language and Environment for Statistical Computing Version 4.4.0, R Foundation for Statistical Computing. Available online: http://www.R-project.org (accessed on 1 February 2024).
33. Foundation for Open Access Statistics, Journal of Statistical Software. Available online: https://www.jstatsoft.org (accessed on 1 June 2024).

34. Jackman, S.; Tahk, A.; Zeileis, A.; Maimone, C.; Fearon, J.; Meers, Z. *Package 'pscl' Version 1.5.9*; Political Science Computational Laboratory; CRAN of R Project; R Foundation for Statistical Computing: Vienna, Austria, 2023.
35. Meyer, P.G.; Cherstvy, A.G.; Seckler, H.; Hering, R.; Blaum, N.; Jeltsch, F.; Metzler, R. Directedeness, correlations, and daily cycles in springbok motion: From data via stochastic models to movement prediction. *Phys. Rev. Res.* **2023**, *5*, 043129. [CrossRef]