A Clustering Method of Highly Dimensional Patent Data Using Bayesian Approach

Sunghae Jun

Department of Statistics, Cheongju University
Cheongju, Chungbuk 360764, Korea

Abstract

Patent data have diversely technological information of any technology field. So, many companies have managed the patent data to build their R&D policy. Patent analysis is an approach to the patent management. Also, patent analysis is an important tool for technology forecasting. Patent clustering is one of the works for patent analysis. In this paper, we propose an efficient clustering method of patent documents. Generally, patent data are consisted of text document. The patent documents have a characteristic of highly dimensional structure. It is difficult to cluster the document data because of their dimensional problem. Therefore, we consider Bayesian approach to solve the problem of high dimensionality. Traditional clustering algorithms were based on similarity or distance measures, but Bayesian clustering used the probability distribution of the data. This idea of Bayesian clustering becomes a solution for the problem in this research. To verify the performance of this study, we will make experiments using retrieved patent documents from the United States Patent and Trademark Office.

Keywords: *Patent Clustering, Bayesian Clustering, Highly Dimensional Problem, Probability Distribution, Bayesian Learning.*

1. Introduction

The applied patent documents have been increased rapidly according to the progress of intellectual property systems. Most governments and companies have applied their results of the developed technologies to the patents in the world. Also, they analyze the patent data for efficient patent management because the patent data have diversely technological information of any technology field. Many companies have managed the patent data to build their R&D policy. Patent analysis is an approach to the patent management. Also, patent analysis is an important tool for technology forecasting. The classification and clustering of patent documents are popular methods of patent analysis. In this paper, we propose a clustering method of patent documents. The patent document data are sparse and have high dimensional data structure. So, it is difficult to cluster the patent documents by traditional clustering methods such as hierarchical and K-means clustering algorithms. To solve the dimensional problem of the patent data, we consider Bayesian clustering in this research. The idea of Bayesian analysis is to combine previous and current information of given domain data [1]. There were so many researches of Bayesian approach [2-3]. Efficient clustering was a popular issue of Bayesian data analysis [4-5]. Most results of Bayesian clustering were model based clustering [6]. This clustering method provided the determination of the number of clusters optimally [7]. So, we will use this approach to select the number of clusters in this paper. Bayesian clustering provides the number of clusters by the posterior values using Bayes' rule. The clustering is to assign multivariate data into a number of clusters by similarity or distance measures [8]. Another clustering approach was introduced by Symons [9]. This method used a probabilistic model by mixtures of multivariate normal distribution. Bayesian hierarchical clustering was one of Bayesian clustering approaches [10]. In this paper, we construct a dendrogram as a clustering result using posterior probability as a distance value. This is one of the results of Bayesian hierarchical clustering. We will also build a log posterior plot for determining the number of clusters. We make experiment for verifying the improved performance of this research using retrieved patent documents from United States Patent and Trademark Office (USPTO) [11]. The topic of retrieved patent documents is social network service (SNS). In our experiment, we will cluster the searched patent data using Bayesian clustering.

2. Background Research

2.1 Clustering Patent Data with High Dimension

Patent document clustering is similar to general document clustering. But, the data type of patent document is different to general document in one instance. That is, patent document has issued date, citation, and family patent information. These are not in general document. So, we need to consider the characteristics of patent document. One of them is high dimensionality problem. We have to transform the patent document data into structured data for quantitative clustering methods. In this process, the high dimensionality problem is occurred. The structured data are consisted of large matrix with documents and terms. The row and column represent document and term respectively. Each value of the matrix is the occurred frequency of a term in each document. Also, most values are 0 because the dimension of the matrix is so high. So, it is difficult to analyze the matrix. To settle this difficulty, we consider Bayesian analysis in this paper. Using the probabilistic approach of Bayesian clustering, we will cluster the patent documents.

2.2 Bayesian Clustering

A general method of Bayesian clustering was based on finite mixture model of probability distribution [12]. This probability model partitions a set of elements to some clusters using predictive distribution [13]. That is, Bayesian clustering is performed by the conditional probability of comparing to the existing clusters. Sometimes, the clustering makes a new cluster by the probability distribution. The elements and clusters are represented by $D=\{d_1, d_2, ..., d_n\}$ and $C=\{c_1, c_2, ..., c_k\}$ respectively. We can define the joint probability distribution of the elements as follow.

$$p(d_1,d_2,...,d_n)=\prod_{i=2}^{n}p(d_i\,|\,d_{i-1},...,d_2,d_1) \quad (1)$$

Where, the probability of $p(d_1, d_2, ..., d_n)$ can be defined as the conditional probability of $p(d_i|d_{i-1}, ..., d_2, d_1)$. Using this predictive distribution, we can assign the elements to existing clusters or new clusters. Markov Chain Monte Carlo (MCMC) method is more advanced Bayesian approach [3]. But, this demands larger computing resource and time cost for analyzing the given data. This research does not consider MCMC method because we have the highly dimensional problem of the document data.

3. A Proposed Method

Bayesian analysis has a powerful method to combine current data information with previous analyzer's knowledge. The combined result produces updated information for optimal decision. This is based on Bayes' rule [14]. The parameter θ of Bayesian model is a random variable [14]. This has a probability distribution. That is, the prior distribution of the parameter $p(\theta)$ is defined subjectively and initially without the information of given data. The other way, likelihood distribution is based on given data x. This is a conditional distribution, $l(x|\theta)$. By multiplying the prior distribution and likelihood function, Bayes' rule computes the posterior distribution $p(\theta|x)$ of the parameter of Bayesian model as follow [15].

$$p(\theta\,|\,x)=\frac{l(x\,|\,\theta)\,p(\theta)}{P(x)} \quad (2)$$

This is the analyzer's updated belief of the parameter of Bayesian model for optimal decision (clustering). To solve the highly dimensional problem of patent document clustering, we use this Bayesian updating approach. Bayes' theorem was updated as the following formula [14].

$$Posterior \propto Prior\times Likelihood \quad (3)$$

The prior has the previous knowledge of given data. Also, the current information of the data is included in the likelihood function. So, we can compute posterior probability by multiplying prior distribution and likelihood function. In this paper, we construct a clustering method for the patent document clustering using Bayesian updating. Bayesian clustering is a probabilistic model on data partition for efficient document clustering. In this paper, we use the hierarchical model of Bayesian clustering for suitable clustering of the patent documents with the problem of high dimensionality. This research constructs a hierarchical dendrogram by posterior probability as a similarity measure. We are able to partition the patent documents to the defined clusters using posterior result of Bayesian updating. Also, we will verify the efficiency of this research by efficient clustering of the patent documents. We have to select the parameter of Bayesian distribution family for the document clustering. The retrieved patent documents are transformed to matrix data by the preprocessing of text mining. The column of this matrix is the extracted term (cluster variable). Each document (observation) is represented by a row of the matrix. This matrix is used for input data of the hierarchical Bayesian clustering in this paper. The distribution family of Bayesian model has Gaussian and Laplace in this model. The proposed method has two steps of input and output as follow.

Input:
(1) Given data, $X=\{x_1, x_2, ..., x_n\}$
(2) Prior distribution, $p(\theta)$
(3) Likelihood function, $l(x|\theta)$

Output:
(1) Posterior distribution, $p(\theta|x)$
(2) Updated parameters of hierarchical Bayesian model
(3) Dendrogram of Bayesian clustering

(Step1) Initialization
(1-1) Setting parameters $(\theta_1, \theta_2, ..., \theta_p)$ of prior distribution

(1-2) Computing likelihood function $l(X|\theta_1, \theta_2, \ldots, \theta_p)$
(1-3) Updating posterior distribution $p(\theta_1, \theta_2, \ldots, \theta_p|X)$
(1-4) Selecting the number of clusters, k
(1-5) Setting $X=\{x_1, x_2, \ldots, x_n\}$

(Step2) Repetition (Clustering), k>1
(2-1) Searching x_i and x_j with input and output with the highest probability
(2-2) Assigning x_i and x_j into x_k

We initialize the parameters and the prior distribution types of Bayesian model in step1. The number of clusters and data are built in this step. That is, all parameters of Bayesian model such as prior distribution are decided in step1. In the step2, we repeat the Bayesian updating for finding proper parameters of final Bayesian model for patent documents assignment. Therefore, we perform the efficient clustering by combining the results of step1 and step2. We also use the result of the dendrogram of the hierarchical Bayesian clustering.

4. Experimental Result

We used the retrieved patent documents from USPTO to verify our improved performance. The patent data of 'social network service (SNS)' were searched for this experiment. We retrieved all applied patent documents in the U.S. up to now. Total 71 patent documents were searched in this research. The following figure shows the number (frequency) of issued patent documents until December 1, 2011. Also, we use R-project as the analytical tool for Bayesian clustering in this experiment [16-17].

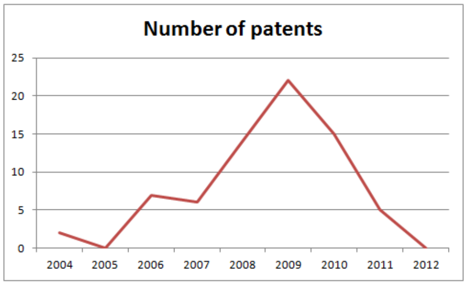

Fig. 1 Number of applied patents by year.

The first patent of the SNS was applied in 2004. We knew the SNS patent documents were increased 2008 through 2009 rapidly. But, the number of applied patents of SNS was decreased after that. So, we concluded that most needed technologies of SNS were developed until 2009. This is considerable for the R&D planning of SNS technologies. Next, we transformed the retrieved patent documents into matrix data. In this paper, we had two steps to construct the matrix data for Bayesian clustering. We constructed the first matrix consisting of retrieved documents and extracted terms. Next, we made second matrix from the first matrix. This matrix had the documents and predicted principal component (PC) scores. All PC scores were computed by principal component analysis (PCA) using the correlation structure of the terms [8]. Next table shows the first (term based matrix) and second (PC score based matrix) data.

Table 1: First and second matrix data

Matrix type	Row	Column
First matrix	71 documents	1559 terms
Second matrix	71 documents	71 PC scores

In this paper, we used the second matrix for input data of Bayesian clustering. The second data had smaller dimension than first data. We identified the result of dendrogram of the hierarchical Bayesian clustering. This dendrogram is shown in next figure.

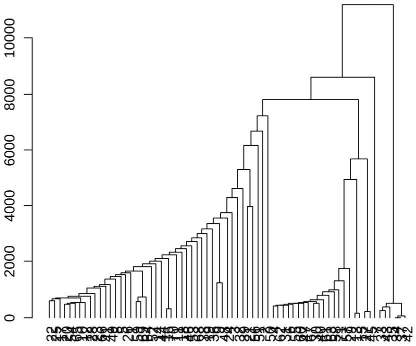

Fig. 2 Dendrogram of hierarchical Bayesian clustering.

The X and Y axes represent patent documents and log posterior values respectively. We could the hierarchical structure of the SNS patent data using the dendrogram. In this experiment, we found the number of clusters for SNS patents clustering from the following figure.

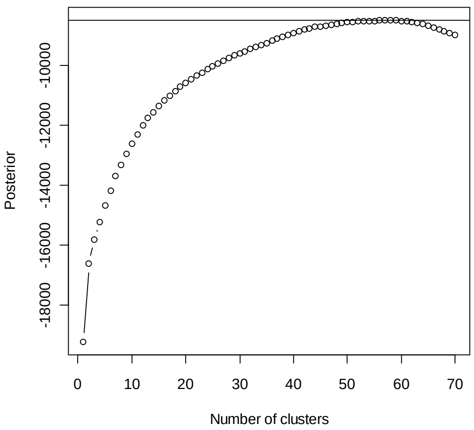

Fig. 3 Number of clusters by posterior.

The result of the hierarchical Bayesian clustering for SNS patent (technology) data gives about 55 clusters. This meant that the developed technologies of SNS were in diverse and wide areas. Next figure represents the importance of clusters variables. In this experiment, we used 71 cluster variables.

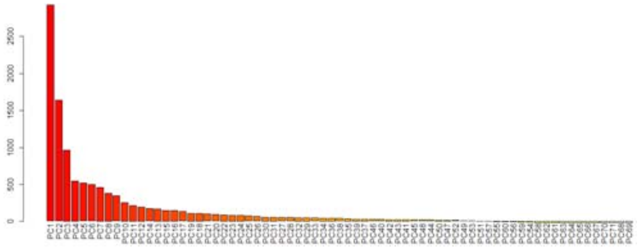

Fig. 4 Importance of cluster variables.

We knew that a few variables affected the clustering result among 71 clustering variables. The largest and smallest important variables were PC1 and PC69 respectively.

5. Conclusions and Future Works

In this paper, we proposed a clustering method of patent documents using the hierarchical Bayesian clustering. General clustering methods such as K-means clustering had some problems for patent documents clustering because the patent data are sparse and have high dimensionality. So, in this research, we considered Bayesian clustering as an efficient method for of the patent document clustering. In this experimental result, we verified the performance of this study. We determined the optimal number of clusters by the log posterior distribution. Also, we assigned the retrieved patent documents using the result of Bayesian clustering. Also, we found the technological trend of SNS technology.

Our future work is to develop more advanced Bayesian clustering model by the MCMC method and Bayesian mixture mode for the patent document clustering.

References

[1] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, Bayesian Data Analysis, Chapman & Hall, 1995.

[2] M. D. Escobar, and M. West, “Bayesian density estimation and inference using mixtures”, Journal of the American Statistical Association, Vol. 90, 1995, pp. 577-588.

[3] G. J. McLachlan, and K. E. Basford, Mixture models: Inference and applications to clustering, Marcel Dekker, 1988.

[4] J. D. Banfield, and A. E. Raftery, “Model-based Gaussian and non-Gaussian clustering”, Biometrics, Vol. 49, 1993, pp. 803-821.

[5] F. A. Quintana, and P. L. Iglesias, “Bayesian Clustering and Product Partition Models”, Journal of the Royal Statistical Society Series B, Vol. 65, 2003, pp. 557-574.

[6] C. Fraley, and A. E. Raftery, “Model-Based Clustering, Discriminant Analysis, and Density Estimation”, Journal of the American Statistical Association, Vol. 97, 2002, pp. 611-631.

[7] C. Fraley, and A. Raftery, “How many clusters? Which clustering methods? Answers via model-based cluster analysis”, Computer Journal, Vol. 41, 1998, pp. 578-588.

[8] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

[9] M. J. Symons, “Clustering criteria and multivariate normal mixtures”, Biometrics, Vol. 37, 1981, pp. 35-43.

[10] K. A. Heller, and Z. Ghahramani, “Bayesian hierarchical clustering”, Proceedings of the 22nd International Conference on Machine Learning, 2005.

[11] United States Patent and Trademark Office (USPTO), www.uspto.gov

[12] J. M. Bernardo, and J. Giron, “A Bayesian approach to cluster analysis”, Questiio, Vol. 12, No. 1, 1988, pp. 97-112.

[13] F. A. Quintana, “A Predictive View of Bayesian Clustering”, Journal of Statistical Planning and Inference, Vol. 136, Iss. 8, 2006, pp. 2407-2429.

[14] S. J. Press, Bayesian Statistics: Principles, Models, and Applicaions, Wiley, 1989.

[15] P. Giudici, Applied Data Mining, Statistical Methods for Business and Industry, Wiley, 2003.

[16] V. P. Nia, R Package ‘bclust’, R-Project, 2011.

[17] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. 2010.

Sunghae Jun is associate professor in the department of statistics, Cheongju University, Korea. He received the B.S., M.S., and Ph.D. degrees in department of statistics from Inha University, Korea, in

1993, 1996. Also he took the doctor's degree in computer science and engineering from Sogang University, Korea, 2007. He worked in NCR as a data mining consultant from 2000 to 2001. He was a visiting scholar in department of statistics, Oklahoma State University, OK, USA from 2009 to 2010. His research fields are machine learning, evolutionary computing, and management of technology.