A Patent Analysis Model Combining Document Clustering and Time Series Analysis

Sang Sung Park[1] and Sunghae Jun[2]

[1] Division of Information Management Engineering, Korea University,
136701 Seoul, Korea
hanyul@korea.ac.kr
[2] Department of Statistics, Cheongju University,
360764 Cheongju, Korea
shjun@cju.ac.kr

Abstract. Patent analysis (PA) is to analyze the data of patent documents such as abstract and the number of issued patents. PA plays a major role in R&D policy because we can forecast the future aspect of a technology by the result of PA. So, most companies make efforts to perform PA for improving their competitiveness. In this paper, we propose new PA model for technology forecasting (TF). We will combine the clustering and predictive results for TF. Using the retrieved patent documents from the United State Patent and Trademark Office, we will make experiment to verify the performance of this research.

Keywords: Patent analysis, Document clustering, Technology forecasting.

1 Introduction

Technology forecasting (TF) is to foresight the technological aspect in future [1-2]. A considerable portion of the R&D plan has been depended on the TF results. Also, patent analysis (PA) plays an important role in the TF process. So, in this paper, we propose a PA method for TF. We use the document clustering and time series analysis and combine these results for constructing efficient TF model. Many researches were published in PA fields [3-5]. Most of them were based on one analytical approach such as clustering, classification, and citation analyses. But, they had some limitations to forecast the future state of a technology because they were depended on only one result of a TF method. So, to reduce this problem, we consider combining two analytical approaches which are patent document clustering and time series model. We use K-means clustering algorithm [6-7] as a patent clustering method and time series regression (TSR) [8] as a time series model. To verify the performance of this research, we perform a case study using the patent documents related to biotechnology from the United State Patent and Trademark Office (USPTO) [9].

2 New PA Model

In this paper, we propose a PA model for TF. This research consists of two analytical methods which are K-means clustering algorithm and TSR model. In other words, we combine the results of two methods for efficient TF. K-means clustering is a clustering method for finding K clusters and assigning all points to each cluster by Euclidean distance measure [10]. TSR is a time series method to model the function of dependent variable Y and independent variable X, where X is time. In this paper, the TSR trend model is defined as follows [8].

$$Y_t = \beta X_t + \varepsilon_t \tag{1}$$

Where Y_t is the issued number of patent documents in period t. X_t and ε_t are the time period t and the error term in time period t respectively. Also, β is a regression parameter of TSR. For the assessment of TRS model, we use the coefficient of determination and probability value (p-value) of the regression parameter.

3 Conclusions

In this paper, we proposed a PA model using document clustering and time series analysis. We used K-means clustering algorithm for patent clustering. Also, we constructed the technological trend model of each cluster using TSR. The aim of this paper was to combine the results of patent clustering and TSR model effectively. In our work, the retrieved patent documents were transformed into structured data using text mining techniques for the quantitative analysis. Using the structured data, we performed our combined model to find emerging area of biotechnology. We determined this area from the results of patent clustering and TSR model. In the clustering result, we decided the technology of the cluster with relatively small size to the emerging technology. Also, we verified its significance by the TSR result.

This research contributes to a TF domain to search the emerging technology. But, our work had a limitation which was to select the emerging technology subjectively. In other words, we need the domain expert's knowledge to define the technology of the clusters. In our future work, we will develop more objective TF approach to find the emerging technology.

Acknowledgements

This work was supported by the Brain Korea Project in 2012.

References

1. Hunt, D., Nguyen, L., Rodgers, M.: Patent Searching Tools & Techniques. Wiley (2007)
2. Roper, A. T., Cunningham, S. W., Porter, A. L., Mason, T. W., Rossini F. A. and Banks J.: Forecasting and Management of Technology. Wiley (2011)
3. Jun, S., Uhm, D.: Technology Forecasting Using Frequency Time Series Model: Bio-Technology Patent Analysis. Journal of Modern Mathematics and Statistics, 4(3), 101--104 (2010)
4. Bengisu, Z., Nekhili, R.: Forecasting emerging technologies with the aid of science and technology databases. Technological Forecasting and Social Change, 73(7), 835--844 (2006)
5. Tseng, Y., Juang, D., Wang, Y., Lin, C.: Text mining for patent map analysis. Proceedings of IACIS Pacific Conference, 1109--1116 (2005)
6. Everitt, B. S., Landau, S., Leese, M.: Cluster Analysis, fourth edition. Apnold (2001)
7. Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kaufmann (2001)
8. Bowerman, B. L., O'Connell, R. T., Koehler, A. B.: Forecasting, Time Series, and Regression, An Applied Approach, Brooks/Cole (2005)
9. United State Patent and Trademark Office (USPTO), www.uspto.gov
10. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Springer (2001)