A Study on News Analysis System for Extracting Issue Items

Park, Young Wook*, Bae, Kuk-Jin**, Choi, Boong-Kee***, Choi,Yunjeong ****, Park, Jin-Woo*****
*,**,***,****Korea Institute of Science and Technology Information, *****Diquest
E-mail : {ywpark, baekj, boongkee, yjchoi}@kisti.re.kr, jwpark@diquest.com

1. Introduction

Trend analysis is important in that we can predict future by finding the people's interest. Jeremy et.al proposed trend analysis method to detect influenza epidemics using search engine query data[1] and Hyeyong et.al analyzed a trend of cultural consumption based on newspaper texts. These papers are related to trend analysis using big data like query data and newspaper texts. We can apply the method to information providing system in business area for finding issue items which could be a clue of new business.

2. Proposed News Analysis System

We proposed news analysis system(NAS) which provides issue items useful to develop new products or a new business service. In this paper items mean a product or a technology that small and medium-sized companies may have a high interest in. Such items are mentioned many times in newspaper articles, besides the frequency are increasing over time.

2.1. Database

Database is a collection of a newspaper articles provided by Naver(www.naver.com). We collect articles in business, society, life/culture and IT/Science section among them. The collected article is over 6 million since September 2011.

2.2. Data Management Process

Data Management Process is composed of a few steps; Document Collecting, Text Mining, Indexing & Searching and Analysis Service.

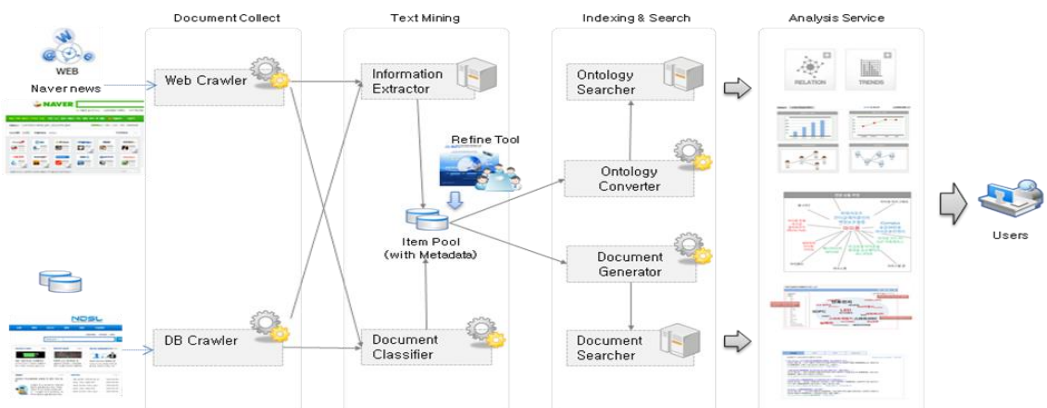

Figure 1. Data Management Process

Firstly Web Crawler collects newspaper articles from Naver and stores in system DB. In Text Mining step, NAS extracts information and classifies articles. Morphme analysis is used in this step. Data is checked by manager before being transfred to next step. In Indexing & Searching step, NAS makes index information to link issue item and newspaper articles. This step costs over 12 hours time to finish, which is a longest task in data management process. Analyzed data is converted to ontology and NAS builds various types of statistical data and related information. The last step, analysis service provides information about issue item to users.

2.3. Statistical Method for Trend Analysis

We applied Mann-Kendall test and t-score to trend analysis. The Mann-Kendall test is a non-parametric test for identifying trends in time series data. Each data value is compared to all subsequent data values. The initial value of the Mann-Kendall statistic is assumed to be 0(*e.g.*, no trend)[3]. If the output value is positive number, trend is positive and vice versa. We use co-occurrence keyword analysis to find a keyword related to issue item. If we count only frequency of occurrence, the high ranked keywords would be most generic terms. Therefore, we focus on co-occurrence between issue item and related keywords, and t-score is suitable in this case. High t-score means high relation between 2 words.

3. Using example

We implemented a news analysis system providing a new business idea to small and medium-sized companies. We build a database composed of newspaper articles and extracted 14,000 items we expect to be issue items. Among them, we chose one keyword, 'mobile service'. And we found 5 keywords related to it using t-score. 'Story', 'Small & medium Business Corporation(SBC)','Smile', 'Library', and 'Smart'. The reason for why those keywords are related to mobile service is easily deduced by reading news articles. It was a newspaper issue that SBC introduced a mobile service.

Figure 2. Using Example of News Analysis System

4. Future Work

The main role of proposed news analysis system is to extract issue items in business area. There can happen a redundancy problem by counting same content articles. We can avoid such problem by counting articles only in same newspaper. However, another difficulty remains, we could not guarantee minimum amount of articles to analyze trend information. We should fix balance point between quality and quantity. By the way, we need a strategy to improve extensive use and convenience

5. References

[1] Jeremy Ginsberg, et.al, “Detecting Influenza Epidemics Using Search Engine Query Data”, Nature, Macmillan, 19 February 2009, pp1012-1015.
[2] Kwon *Hyeyoung Kim*, et.al, “A Trend Analysis of Cultural Consumption Based on Newspaper Texts”, Information", Journal of KIISE : Software and Application, vol.39, no.3, pp.244-251 (in Korean)
[3] *HydroGeoLogic, In.," Annual Groundwater Monitoring Report – Former Fort Ord, California", 18 July 2005, Appendix D*