ICIC Express Letters
Part B: Applications
Volume **7**, Number **10**, October **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **2211–2216**

A METHOD FOR EXTRACTING ORGANIZED INFORMATION FROM ONLINE PRODUCT REVIEWS BASED ON TEXT MINING

Jooyoung Kim and Dongsoo Kim*

Department of Industrial and Information Systems Engineering
Soongsil University
369 Sangdo-Ro, Dongjak-Gu, Seoul 06978, Korea
whytimesgone@ssu.ac.kr; *Corresponding author: dskim@ssu.ac.kr

Received March 2016; accepted June 2016

Abstract. *Currently, the Internet which plays a role as a platform enables customers to produce, modify, and distribute the data. Because of this characteristic, the amount of Internet data has increased exponentially. Unstructured data such as natural language format data occupy most of the Internet data. And one of the unstructured data called personal online product review is significant data for both a company and potential customers. From the perspective of the company, it can acquire insights on the marketing strategy or product renewal direction by analyzing the product review data. From the perspective of the potential customers, they can obtain comprehensive assessment results by organizing the entire review data. In order to extract organized review information, the process of collecting, storing, preprocessing, analyzing data, and drawing conclusions is required. We propose a new approach to extracting useful information from product review data based on text mining. In addition, we present an application case of cosmetic products and verify the benefits of the proposed approach.*

Keywords: Text mining, Discovery data mining, Association rules, Product reviews

1. **Introduction.** As arriving at the Web 2.0 era, Internet users are actively involved in producing information and knowledge and sharing a production with others in an age of the open Internet. This means that Internet users can easily provide the information and contents by themselves [6]. Because of these characteristics, the amount of the data on the Internet has increased enough to be called the deluge of information. And most of the companies are trying to extensively apply these Internet data into the data mining for comprehending current state of the business and figuring out insights on customers. Typically, consumers' product preferences can be estimated by using online or paper-and-pencil based surveys for acquiring feedbacks from customers. However, this type of preference elicitation can easily become expensive in terms of time, and the quality of the data resulting from surveys directly depends on the willingness of the respondents who participate in the study [1]. In order to reduce these inefficiencies, companies structure and model the dataset for verifying the association relationship among various attributes, detecting the abnormal cases, and catching the general reputation by the data on the Internet. The company makes decisions referring to the conclusions based on these activities.

One of the kinds of the unstructured data called personal online product reviews provides useful information for the company that produces the relevant products and others who are interested in the products under the review category. However, most of the reviews in the Internet environment are drawn up as the unstructured text format like the natural language format. This means that it is very difficult to extract significant information from the unstructured data without applying the text mining technology. And also, there are massive reviews generated by Internet users' active or passive participation. In order to find out valuable information in need, it is firstly required to collect the unstructured review data and to process the text analysis skills.

Therefore, we introduce the review text mining methodology to apply the natural language processing technology to the unstructured text data like online product reviews in order to analyze the structured data by using R programming. Furthermore, we introduce the data mining technology to derive the organized information by applying the data mining methods to the column's attributes and each element belonging to the product review dataset extracted by the text mining technology.

The remainder of this paper is organized as follows. Section 2 explains related work including the text mining framework and data mining technology. Section 3 describes the proposal for the review text mining framework and presents detailed instructions about each of the steps. Section 4 deals with the application case of performing the review text mining. Finally, Section 5 offers conclusions.

2. **Related Work.** Recently, the volume of information is growing rapidly, while opportunities to expand insights by combining data are accelerating [8]. And social media generates the amount of the unstructured data more than the amount of the structured data in the big data environment. Therefore, it is expected that the unstructured data is more valuable than the structured data. In the same vein, word-of-mouth marketing using the online unstructured review data which have been individually evaluated has become popular. The word-of-mouth marketing has a great influence on both potential customers who may have an interest in the product and the company which has a necessity of an overall evaluation and comprehending the merit and demerit of it [4]. However, lots of dispersed online review data do not enable us to acquire useful information in its own format. Also, reading all of the review data makes potential customers waste their time. Due to this, potential customers decide to purchase the product after reading many popular or latest reviews [7]. Consequently, it is necessary for the discovery of meaningful and rule-based information not only to enumerate the data which were calculated by the simple gross intelligence but also to arrange review data in a very appropriate manner [5]. In order to filter review data, we should apply text mining to the unstructured review data.

Text mining refers to the process of discovering interesting and non-trivial patterns or knowledge from text format information [3]. Each human language not only has the lexical and grammatical uniqueness but also is often hard to define the consistent rules because of the various ranges of expression shapes. And text mining technology is based on the natural language processing skills. Natural language processing technology is used to analyze the language which is expressed in letters of human languages and to understand the structure and meaning.

Most existing studies related to the Korean text mining verified a hypothesis by using the text data extracted from specific domain [9-11]. Lots of studies handle a partial process or a specific industry of the text mining application domain [12]. Therefore, suggesting a comprehensive method capable of continuously analyzing output is necessary for extracting meaningful review information in the deluge of the information.

In plain language, text mining provides mining skills that extract valuable and significant information from the unstructured text data. Because the most natural form of representing information is text format, text mining can be used for the multiple purposes.

Figure 1 shows the process of drawing meaningful patterns from corpus using text mining. This methodology can extract the keywords which represent the product and express the relationship between the words. As the Internet users are able to draw the significant information from rearranging the keywords, it is essential to extract the words from

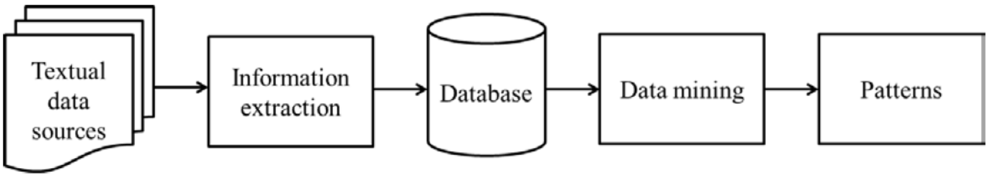

Figure 1. Process of extracting patterns using text mining

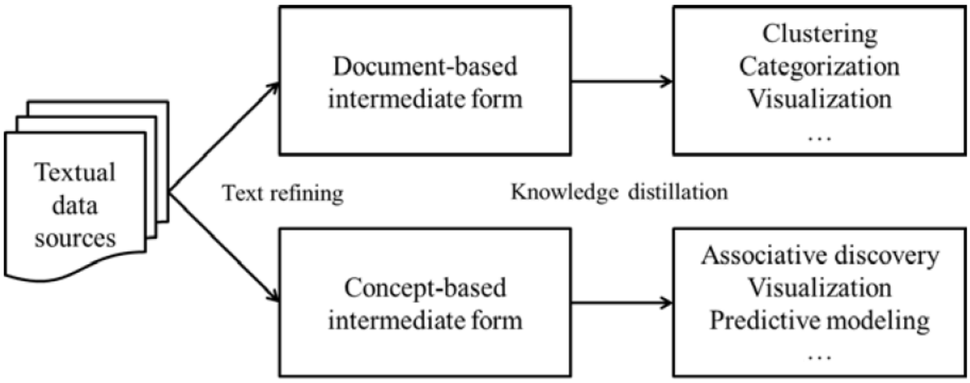

Figure 2. Text mining framework on the basis of intermediate forms (IFs)

collected text data by using the natural language processing in the step of the information extraction. Then, these refined data are stored in the database in order to apply the various data mining technique. And the pattern extracted from the mining presents the correlation between other keywords. Therefore, users can obtain more significant results than simply searched results. When computers can understand the descriptive information with which humans use languages, a great deal of language resources, and complicated statistical and regular algorithms should be applied at this point.

Figure 2 shows a text mining framework consisting of two intermediate forms [2]. A document-based IF's each entity represents a document. And a concept-based IF's each entity is an object in a specific domain. This category contains association discovery, predictive modeling, and so on. In this study, we used the association discovery in the phase of analyzing the keywords correlation.

3. **Proposal for the Review Text Mining.** This section presents the research framework for analyzing product reviews created by customers and extracting meaningful and organized information from the reviews. Figure 3 shows the process of the proposed framework, which consists of collecting data, processing the natural language, extracting the keywords and sorting out data and summary.

The framework above is to make architecture review text mining method. According to the proposed framework, we can extract two kinds of review information. The first one is comprehensive review information and the other one is the related review information based on the user's characteristics. With the comprehensive review information, we can analyze the correlation between review features. For example, there are effective values in function compared to the price. If we review a specific product based on the user's characteristics, we can combine customer information and product reviews, and organize the customer's overall reviews which have similar features. However, in this paper, we focus on the pre-process of review process step and extracting keywords and correlation step.

In advance, it is possible to collect a customer profile such as product review, ID, age, and skin type of the customer in public using the web-crawling method. Collected data can

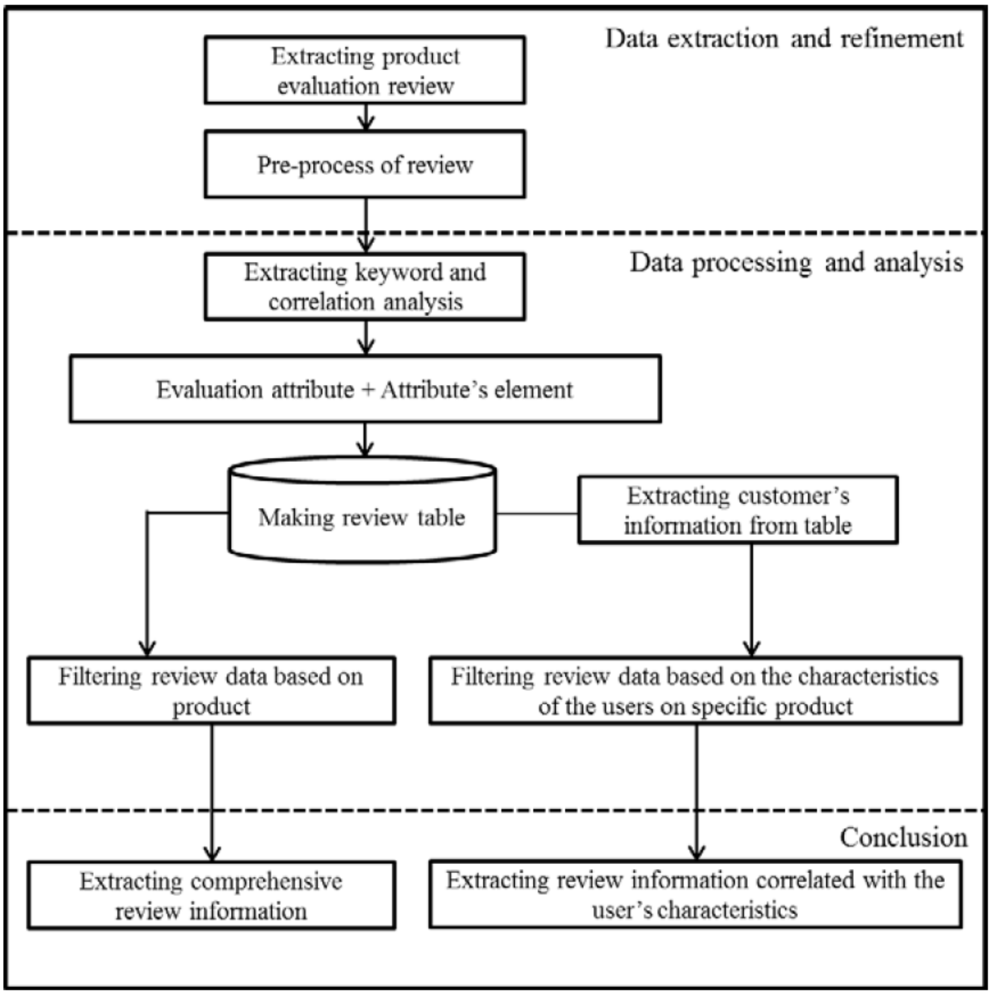
Data extraction and refinement
Extracting product evaluation review
Pre-process of review
Data processing and analysis
Extracting keyword and correlation analysis
Evaluation attribute + Attribute's element
Making review table
Extracting customer's information from table
Filtering review data based on product
Filtering review data based on the characteristics of the users on specific product
Conclusion
Extracting comprehensive review information
Extracting review information correlated with the user's characteristics

Figure 3. Proposed framework for review text mining

be saved in the form of a 2-dimensional database table, and in the form of text file in order to analyze corpus. In the pre-processing stage, firstly, we prepare extracted corpus from the collected product review data and separate each sentence based on the morpheme. At the first time of extracting words, secondly, we use KoNLP package in R and extract the words. Thirdly, we use try-catch function to find out the words not recognized by the data dictionary. Fourthly, we eliminate the useless stopwords and add the unrecognized words which do not exist in the dictionary but necessary for the analysis. Finally, we repeat the third and fourth steps in order to increase the dictionary performance.

In the extracting keywords and correlation analysis stage, in order to select the keyword, we used TF-IDF (Term Frequency – Inverse Document Frequency) method and the frequency of words [13]. TF-IDF is intended to reflect how important a word is to represent the document among the corpus. Through TF-IDF formula, we can estimate the word's degree of representing the feature of the document.

At the end of data extraction and refinement phase, there is data processing and analysis. Here, we can extract data dictionary definition as the pre-stage of analyzing the relationship between the keywords. After finishing data dictionary definition phase, in specific product reviews, we can analyze the relationship between keywords extracted by similar users, understand the relationship between words and cluster the similar words.

After converting the unstructured data into the structured data, we draw organized review information by applying statistical methods, and extract meaningful patterns based on the rules.

4. **Application Case of Review Text Mining.** This section presents an application case of the review text mining. We used cosmetic product reviews to find out how customers think about the product. After pre-process of review step, we extracted 25,361 words about one product. From the words, we extracted the keywords which have high numeric value of representing the feature of the document. And we applied the subset association rule to the extracted keywords by using defined parameters. Finally, we recombined and organized the mining results.

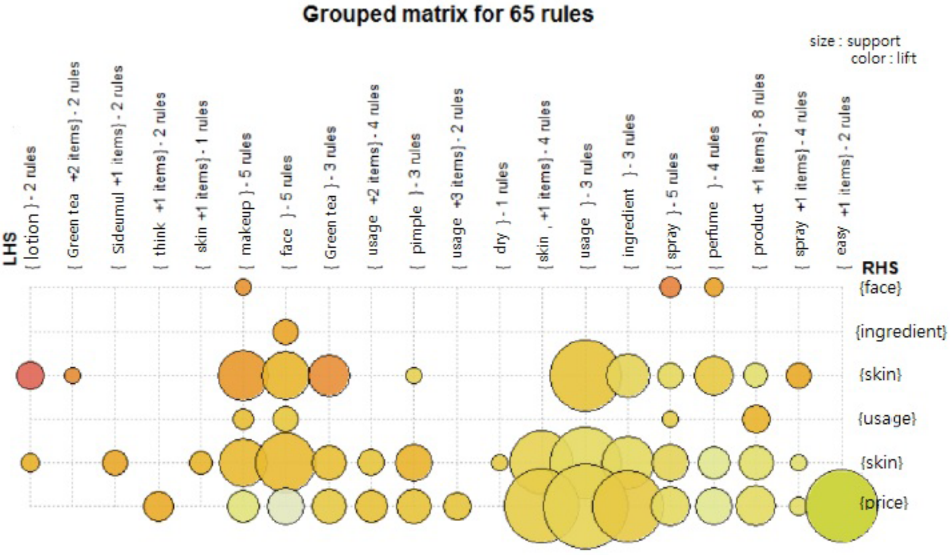

Figure 4. The results of applying review text mining

Figure 4 illustrates the result of applying review text mining. In this figure, left hand side (LHS) means the condition of the association rule. Currently, we confined its maximum count to three in order to focus on each word's correlation. Right hand side (RHS) means the rule's result. The sizes of the circle indicate support value of each rule and the darkness of the color means lift value.

5. **Conclusions.** The Internet users can easily search for lots of data on the Internet. And they can produce Internet information such as online digital product reviews and vast pieces of information on the Internet with ease. However, it is required to use text mining for filtering the unstructured data in order to find the comprehensive meaning among a great amount of data. Aside from the general information through overall data, categorizing and grouping the data is necessary and helpful for the companies and consumers that want to find the specific information focused on special conditions. This research suggested a new method that extracts meaningful review information from a vast amount of data using review text mining.

However, this research has some limitations. First, it is difficult to verify the accuracy of every customer's review. Shortly speaking, it is very difficult to distinguish the fabricated reviews written by unethical reviewers from the significant reviews by the real users. Second, the negative words which are relatively less than the positive words were not shown up in the results because of the method of extracting keywords.

In the near future, we are going to design and implement a detailed algorithm proposed in this research and propose a way to overcome the above mentioned limitations.

Acknowledgement. This research was supported by the Ministry of Trade, Industry and Energy (MOTIE), Korea, through the Education Program for Creative and Industrial Convergence (Grant Number N0000717).

REFERENCES

[1] R. Decker and M. Trusov, Estimating aggregate consumer preferences from online product reviews, *International Journal of Research in Marketing*, vol.27, no.4, pp.293-307, 2010.

[2] A. H. Tan, Text mining: The state of the art and the challenges, *Proc. of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, vol.8, pp.65-70, 1999.

[3] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From data mining to knowledge discovery in databases, *AI Magazine*, vol.17, no.3, pp.1-37, 1996.

[4] A. Kangale, S. K. Kumar, M. A. Naeem, M. Williams and M. K. Tiwari, Mining consumer reviews to generate ratings of different product attributes while producing feature-based review-summary, *International Journal of Systems Science*, pp.1-15, 2015.

[5] M. Collins, Head-driven statistical models for natural language parsing, *Computational Linguistics*, vol.29, no.4, pp.589-637, 2003.

[6] O. Tim, What is Web 2.0: Design patterns and business models for the next generation of software, *Communications & Strategies*, no.1, pp.17-37, 2007.

[7] N. Ma, E. P. Lim, V. A. Nguyen, A. Sun and H. Liu, Trust relationship prediction using online product review data, *Proc. of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management*, pp.47-54, 2009.

[8] McKinsey Global Institute, *Three Keys to Building a Data-Driven Strategy*, http://www.mckinsey.com/business-functions/business-technology/our-insights/three-keys-to-building-a-data-driven-strategy.

[9] S. K. Kang, H. Yu and Y. J. Lee, Analyzing disaster response terminologies by text mining and social network analysis, *Information Systems Review*, vol.18, no.1, pp.141-155, 2016.

[10] S. G. Cho, J. H. Cho and S. B. Kim, Discovering meaningful trends in the inaugural addresses of United States presidents via text mining, *Journal of the Korean Institute of Industrial Engineers*, vol.41, no.5, 2015.

[11] J. S. Han and J. H. Yoon, Activation strategies of the 20th BIFF using social big data text mining analysis, *Journal of the Tourism Sciences Society of Korea*, vol.40, no.1, pp.133-145, 2016.

[12] S. G. Kim, H. J. Cho and J. Y. Kang, The status of using text mining in academic research and analysis methods, *Journal of Information Technology and Architecture*, vol.13, no.2, 2016.

[13] Wikipedia, *TF-IDF*, https://en.wikipedia.org/wiki/Tf%E2%80%93idf.