ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

텍스트 마이닝 기반의 이슈 관련 R&D 키워드 패키징 방법론☆

Methodology for Issue-related R&D Keywords Packaging Using Text Mining

현 윤 진[1] 윌 리 엄[1] 김 남 규[2*]
Yoonjin Hyun William Wong Xiu Shun Namgyu Kim

요 약

빅데이터 기술에 대한 관심이 급증함에 따라, 소셜 미디어를 통해 유통되는 방대한 양의 비정형 데이터를 분석하고자 하는 시도가 활발히 이루어지고 있다. 이에 따라서 텍스트 형태의 비정형 데이터 분석을 통해 의미 있는 정보를 찾고자 하는 시도가 비즈니스 영역뿐 아니라, 정치, 경제, 문화 등 다양한 영역에서 이루어지고 있다. 특히 최근에는 여러 현안 및 이슈들을 발굴하여 이를 의사결정에 활용하고자 하는 시도가 활발히 이루어지고 있다. 이처럼 빅데이터 분석을 통해 국가현안이나 이슈를 발굴하고자 하는 시도가 꾸준히 이루어져왔음에도 불구하고, 국가현안 및 이슈로부터 이와 관련된 R&D 문서를 효율적으로 제공하는 방안은 마련되지 않고 있다. 이는 사용자들이 인식하는 현안 키워드와 실제 사용되는 R&D 키워드 사이의 이질성이 존재하기 때문이다. 따라서 현안 및 R&D 키워드간의 이질성을 극복하기 위한 중간 장치가 필요하며, 이 중간 장치를 통해 각 현안 키워드와 R&D 키워드간에 적절한 대응이 이루어져야 한다. 이를 위해 본 연구에서는 (1) 현안 키워드 추출을 위한 하이브리드 방법론, (2) 현안 대응 R&D 정보 패키징 방법론, 그리고 (3) R&D 관점에서의 연관 현안 네트워크 구축 방법론의 총 세 가지 방법론을 제안한다. 제안하는 방법론은 텍스트 마이닝, 소셜네트워크 분석, 그리고 연관 규칙 마이닝 등의 데이터 분석 기법들을 활용하여 수행하였으며, 그 결과, (1)에 의한 키워드 보강률은 42.8%로 나타났으며, (2)의 경우, 현안 키워드와 R&D 키워드간 다수의 연관 규칙이 나타났다. (3)의 경우는 현재 진행 중에 있으며, 향후 가시적 성과를 낼 수 있을 것으로 예상된다.

☞ 주제어 : 연관 규칙 마이닝, 키워드 매칭, 소셜네트워크 분석, 텍스트 마이닝, 토픽 분석

ABSTRACT

Considerable research efforts are being directed towards analyzing unstructured data such as text files and log files using commercial and noncommercial analytical tools. In particular, researchers are trying to extract meaningful knowledge through text mining in not only business but also many other areas such as politics, economics, and cultural studies. For instance, several studies have examined national pending issues by analyzing large volumes of text on various social issues. However, it is difficult to provide successful information services that can identify R&D documents on specific national pending issues. While users may specify certain keywords relating to national pending issues, they usually fail to retrieve appropriate R&D information primarily due to discrepancies between these terms and the corresponding terms actually used in the R&D documents. Thus, we need an intermediate logic to overcome these discrepancies, also to identify and package appropriate R&D information on specific national pending issues. To address this requirement, three methodologies are proposed in this study—a hybrid methodology for extracting and integrating keywords pertaining to national pending issues, a methodology for packaging R&D information that corresponds to national pending issues, and a methodology for constructing an associative issue network based on relevant R&D information. Data analysis techniques such as text mining, social network analysis, and association rules mining are utilized for establishing these methodologies. As the experiment result, the keyword enhancement rate by the proposed integration methodology reveals to be about 42.8%. For the second objective, three key analyses were conducted and a number of association rules between national pending issue keywords and R&D keywords were derived. The experiment regarding to the third objective, which is issue clustering based on R&D keywords is still in progress and expected to give tangible results in the future.

☞ keyword : Association Rules Mining; Keyword Matching; Social Network Analysis; Text Mining; Topic Analysis

[1] Graduate School of Business IT, Kookmin University, Seoul, 136-702, Korea.
[2] Associate Professor, School of MIS, Kookmin University, Seoul, 136-702, Korea.
* Corresponding author (ngkim@kookmin.ac.kr)

[Received 7 April 2014, Reviewed 11 April 2014, Accepted 3 June, 2014]
☆ A preliminary version of this paper was presented at ICONI 2013 and was selected as an outstanding paper.

1. INTRODUCTION

The volume of unstructured data generated by social media has been increasing rapidly in recent times. Majority of this data cannot be handled effectively using traditional data analysis methodologies. Particularly, the volume of unstructured text data is huge and complex and a lot of attempts have been made to analyze this data in different areas such as politics, economics, and business [1, 2, 3]. For example, several studies have examined national issues by analyzing large volumes of text related to various social issues.

In traditional approaches, the selection of national pending issues is made by a few policy makers in a top-down manner. However, these approaches have certain limitations as they may fail to reflect rapidly changing social issues. To overcome these limitations, data-driven issue keyword selection approaches have been developed by researchers. Unfortunately, data-driven methods cannot distinguish between keywords pertaining to national pending issues (national pending issue keywords) and simple gossip keywords. Thus, the first goal of this study is to establish a hybrid approach for extracting and integrating national issue keywords in order to overcome the limitations of the top-down and data-driven approaches.

Further, given the discrepancies between issue keywords recognized by users and R&D keywords pertaining to the issues, it is very difficult to provide satisfactory information services that can identify and retrieve appropriate R&D documents related to specific national issues. Thus, a solution that overcomes these discrepancies is required to identify appropriate R&D information related to specific national issues. Therefore, the second goal of this study is to establish a methodology for packaging R&D information that corresponds to national issues.

Moreover, most national pending issues contain problems that need to be resolved using R&D information. Consequently, several attempts have been made to determine the associations among these issues in order to obtain the relevant R&D information more easily. Most of these attempts assume that the associations can be determined by simply investigating the simultaneous occurrence of these issues in the news, articles, and technical documents. Unfortunately, if the issues do not co-occur frequently, co-occurrence-based approaches may neglect some associations even though they have many R&D keywords in common. Thus, the third goal of this study is to establish a methodology for constructing an issue network based on common R&D keywords.

2. Related Works

2.1 Text Mining and Topic Analysis

Text is the most widely used means for exchanging information and expressing intentions in the real world [4]. Thus, several researchers have attempted to discover meaningful knowledge through text analysis. Topic analysis, in particular, is one of the most representative methods for discovering core keywords from a large amount of documents. The key concepts of topic analysis are vector space model [1, 5, 6] and TF-IDF. Once the keywords are refined based on specific conditions, the co-occurrence patterns for each keyword are calculated based on TF-IDF weights. There is also a study on improving the accuracy of text mining [7, 8].

2.2 Association Analysis and Social Network Analysis

Association rules mining [9] is an analysis technique that looks for patterns in co-occurrence of multiple data. The results of this analysis can be expressed in the form of a number of rules based on frequency and co-occurrence probability of the target data. Association analysis has been utilized in various fields [10]. Further, there are studies that use both association analysis and social network analysis [11, 12, 13, 14, 15, 16]. Social network analysis is a quantitative analysis technique that identifies the characteristics of the connection structure and connection state of an object in a group through visual representation [17, 18, 19, 20, 21].

2.3 Keywords Similarity Analysis

The F-score is typically used for measuring the similarity between documents or terms. It is obtained by calculating the harmonic mean of the recall and precision. In document searching applications, recall can defined as the proportion of documents that have been detected to all the relevant

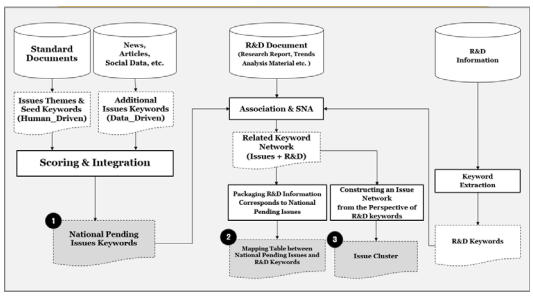
documents. On the contrary, precision means the proportion of appropriate documents to all retrieved documents. In this manner, F-Score is widely used for performance evaluation of information retrieval and also for similarity analysis among multiple documents. [22].

2.4 National R&D Information Service

There are several studies on integrating and managing research outcomes [23, 24, 25, 26, 27] for the purpose of establishing an efficient planning and management system for national R&D projects, developing a seamless information distribution system among various agents of national R&D projects, and building efficient and effective R&D support systems.

3. Methodology for Packaging R&D Information on National Pending Issues

3.1 Research Overview



(Figure 1) Research Overview

Our study has three specific components (see Figure 1). To make it clear, we used issue as national pending issue for the following explanation. The first component generates a list of issue keywords by integrating standard policy documents and additional issue keywords obtained through data analysis. The second component maps the R&D keywords to relevant issue keywords. The final component constructs an issue network based on the common R&D keywords. In Figure 1, the cylindrical shapes indicate the source data, whereas the rectangular shapes represent the analytical process. The boxes outlined with dotted lines represent the outputs.

Next, we provide detailed explanations of each part of the process.

3.2 Hybrid Approach for Extracting and Integrating National Pending Issue Keywords

Traditionally, a small number of experts select national pending issues in a top-down manner. However, this method cannot address all the issues of a complex society. Further, it hardly reflects rapidly changing social issues. To overcome these limitations, data-driven keyword selection methods have been studied by various data scientists as alternatives. However, these methods cannot distinguish between keywords pertaining to issue keywords and general gossip keywords. Therefore, in this study, we propose a hybrid approach for selecting and integrating issue keywords in order to overcome the limitations of the top-down and data-driven approaches. The proposed approach is designed such that it can be applied to one of the core modules of the methodology for packaging R&D information services regarding issue.

In the first stage of the process, issue themes and initial seed keywords (IH) are extracted manually from the standard documents pertaining to issue. To append additional issue keywords that might not be explained in standard documents, additional issue keywords (ID) are obtained from market documents such as news articles, columns, discussions, and social media using a data-driven method. Topic analysis is used to capture additional issue keywords from the market documents.

To integrate IH and ID, we developed the WF-Score as a novel similarity measure by modifying the traditional F-Score. In the expression below, #key (A) indicates the number of keywords in a set A, while C is the number of common keywords between IH and ID. For a set of common keywords between IH and ID, mi is the topic weight of the i-th keyword in ID.

If we use the traditional F-Score method, Precision is calculated by the ratio of the number of common keywords in IH and ID to the number of keywords in ID, while Recall is calculated by the ratio of the number of common keywords in IH and ID to the number of keywords in IH. F-Score is the harmonic mean of Precision and Recall. In traditional

F-Score, it considers the number of common keywords in both set. But in our WF-Score, we measure the weighted sum of the degree of each word contribution in a common set, instead of simple count of them. The expressions are as follows.

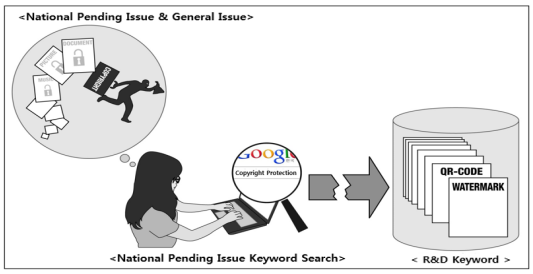
$$\text{Precision}(k) = \frac{\sum_{i=1}^{c} m_i}{\#key(ID_k)}$$

$$\text{Recall}(j) = \frac{\sum_{i=1}^{c} m_i}{\#key(IH_j)}$$

$$\text{WF-Score}(j, k) = \frac{2 \times \sum_{i=1}^{c} m_i}{\#key(IH_j) + \#key(ID_k)}$$

3.3 Packaging R&D Keywords Related to National Pending Issues

Many users attempt to obtain detailed information about issue through various media and press releases. Some users want to acquire more special documents such as technical and R&D reports in order to obtain more professional and systematic information. However, if users are familiar only with issue keywords but not with R&D keywords, then it is difficult for them to obtain appropriate R&D documents through keyword search. The main reason for this phenomenon is the heterogeneity between the professional keywords used in the R&D field and the general issue keywords recognized by the user (see Figure 2).

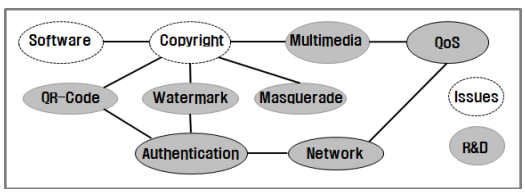


(Figure 2) Heterogeneity between Issue Keywords and R&D Keywords

Figure 2 illustrates an example of a user searching for data related to a copyright infringement case by using the keyword “copyright protection.” If the user is not familiar with R&D keywords such as “watermark” and “QR-Code,” he/she will generally use familiar issue keywords such as “copyright protection” for the initial search query to obtain the necessary technical information. Thus, if the user wants to search for information using issue-related keywords, there is a possibility of not only rigorous trial and error but also failure, as even after numerous attempts, the search may not always yield the appropriate R&D documents. All these limitations demonstrate the heterogeneity between user recognized issue keywords and R&D keywords. To handle this heterogeneity, an intermediate layer is required that connects each issue keyword and the related R&D keyword. For example, associating the R&D keywords “watermark” and “QR-Code” to the issue keyword “copyright protection” makes it possible to provide the appropriate R&D documents corresponding to the issue keyword entered by the user.

In the second stage of the process in Figure 1, we construct a network for packaging R&D keywords pertaining to issue. In the traditional data mining approaches, association analysis concentrates on co-occurrence between keywords in the same dimension [9]. However, in this study, the association rules comprise keywords from two heterogeneous dimensions—issue keywords and R&D keywords. The association rules and their support values are used as inputs to construct a keyword network in the next step [28]. It is possible to identify a set of R&D keywords for each issue by analyzing this two-mode network. To identify the issues that correspond to the R&D keywords, each attribute of the node needs to be assigned to a national pending issue or R&D.

Next, the issues that correspond to the R&D keywords are mapped based on the shortest distance and the number of paths between the nodes. The shortest distance between the two nodes represents the strongest connection between them. In the analysis based on the number of paths, the presence of many paths between two nodes suggests greater reliability of the relationship between them. This study presents a mapping table of R&D keywords that are closely interrelated with each issue, which reflects these two important points.

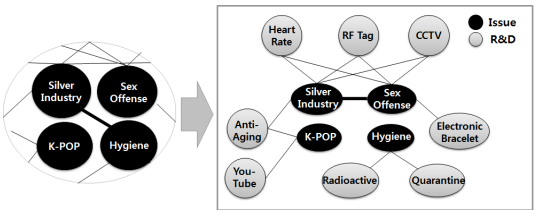


(Figure 3) Two Mode Network Comprising Issue Keywords and R&D Keywords

Figure 3 illustrates an example where R&D keywords related to the issue “copyright” are determined. We set two rules for example: (1) the distance between two nodes must be less than three and (2) the maximum flow between two nodes must be greater than one. The R&D keywords that satisfy these two rules are “QR-Code,” “watermark,” “authentication,” “multimedia,” and “QoS.”

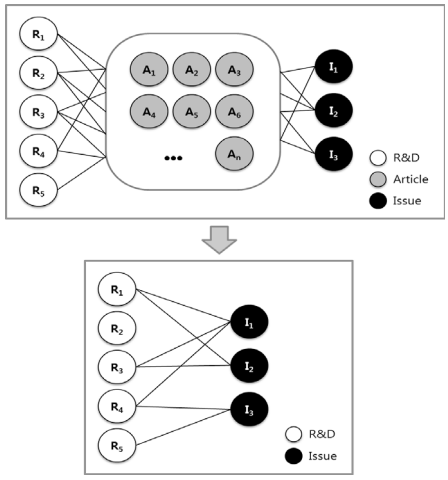
3.4 Methodology for Constructing an Issue Network Based on Common R&D Keywords

In the last stage, the related issues are clustered based on the common R&D keywords. Traditionally, most of the related keywords can be determined by simply investigating the co-occurrence of issues in various documents. In the example shown in Figure 4, what are the issues related to the “Silver Industry”? The Silver Industry includes businesses that focus on products and services for senior citizens. If we examine the left-hand side of the figure that considers only co-occurrence of issues, the terms “Silver Industry” and “Hygiene” may be regarded as strongly related keywords. However, according to the right-hand side of the figure that considers both issues and their related R&D keywords, “Silver Industry” and “Sex offense” share three R&D keywords and have high structural equivalence in terms of R&D. From an R&D perspective, in this way, it is possible to extract the mutually related issues based on structural equivalence. This approach may help to enhance the shareability and reusability of R&D technology.

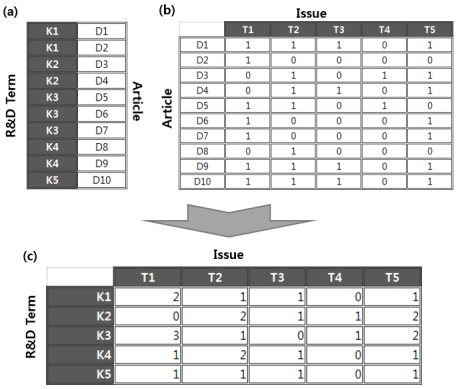


(Figure 4) Discovering Related Issues Using Structural Equivalence

Since the purpose of this stage is to achieve issue clustering from an R&D perspective, it essential to first construct the R&D lexicon based on R&D documents such as patent information or research reports. Thereafter, based on the terms listed in the R&D lexicon, the frequencies of the terms that appear in the news articles are recorded. Next, a network between the news articles and R&D keywords is constructed based on the recorded result. Topic analysis is simultaneously performed using entire articles as target data and a network between the articles and issue is constructed. Next, a network between the issue and R&D keywords is derived by merging the issue/articles network and the articles/R&D keywords network. The network merging process is shown in Figure 5. By using the example in Figure 6(a), it shows the corresponding relationship between R&D and article. Further, Figure 6(b) shows that the matrix between article and issue, which the issue refers to the topic generated through the topic analysis. Lastly, the matrix between R&D and issue is constructed through the network merging process which shown in Figure 6(c).



(Figure 5) Example of Network Merging



(a)	
K1	D1
K1	D2
K2	D3
K2	D4
K3	D5
K3	D6
K3	D7
K4	D8
K4	D9
K5	D10

(b) Issue	T1	T2	T3	T4	T5
D1	1	1	1	0	1
D2	1	0	0	0	0
D3	0	1	0	1	1
D4	0	1	1	0	1
D5	1	1	0	1	0
D6	1	0	0	0	1
D7	1	0	0	0	1
D8	0	1	0	0	0
D9	1	1	1	0	1
D10	1	1	1	0	1

(c) Issue	T1	T2	T3	T4	T5
K1	2	1	1	0	1
K2	0	2	1	1	2
K3	3	1	0	1	2
K4	1	2	1	0	1
K5	1	1	1	0	1

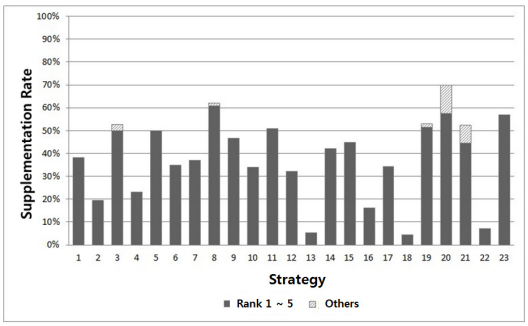
(Figure 6) Example of Matrix (Network Merging)

4. Experiments

4.1 Experiment: Extracting and Integrating National Pending Issue Keywords

In this study, 140 issue themes and 1,148 seed keywords were extracted from standard documents published by the Korean government. Further, additional topics were obtained from 11,000 documents through topic analysis using SAS Enterprise Miner. The final list of integrated issue keywords consisted of 1,148 keywords from standard documents and 1,212 keywords from other documents such as policy reports, news articles, discussions, and columns.

While it is expected that the proposed methodology will augment the number of issue keywords, the degree of enhancement may differ depending on the issue. An analysis of the degree of enhancement of keywords is performed for each of the 140 themes. In the experiment, 1,148 issue keywords and 1,212 general issue keywords were used. Among the 1,212 general issue keywords, 858 keywords were involved in the integrated list and others were ignored. Thus, the keyword enhancement rate of all the national pending issues is about 42.8% (858/2006). Figure 7 shows keyword enhancement rate of each of 23 strategy unit for 140 national pending issues.



(Figure 7) Keyword Enhancement Rate of the Proposed Integration Methodology

Two major results are presented in Figure 7. First, the graph (gray and dark) indicates the results of keyword enhancement for all the national pending issues that matched with each of the general issues. Second, the dark part of the graph is the result of keyword enhancement based on the top five highest matching rates. As seen in Figure 7, these two results are quite similar. In other words, it was found that considering only top five matching rates of national pending issues matching does not weaken the performance of the overall methodology. Further, it shows that the keyword enhancement rate is significantly different for each strategy. In particular, strategies 13, 18, and 22 exhibited a very low enhancement rate, while strategies 8, 20, and 23 exhibited a relatively high enhancement rate. A high enhancement rate was observed for issues pertaining to welfare, unity, and integrity of the government. Strategies 13, 18, and 22 cover topics such as culture, decentralization, and government 3.0. However, it does not mean that the low enhancement rate implies inappropriateness of the selection of national pending issues. On the contrary, it should be noticed that there exist some issues that are considered extremely important from the perspective of policymakers but do not have much public interests.

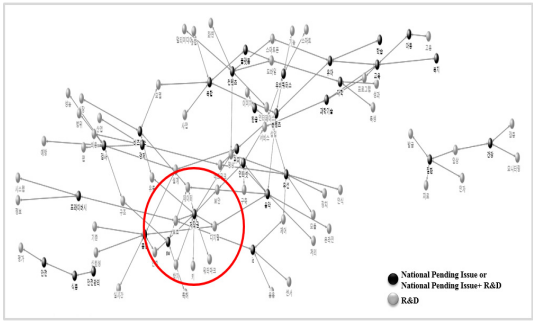
4.2 Experiment: Packaging R&D Keywords Related to National Pending Issues

The documents that contain both issue keywords and R&D keywords were collected for this study. Among the collected documents, documents provided by the NTIS (National Technical Information Service) were selected for the analysis. From 28,060 research reports provided by NTIS, only the main part of the research reports that were published from 2010 – 2012 are considered. Among the research reports, 100 reports whose summaries contain both issue keywords and R&D keywords were selected as the target for the proposed methodology. The extraction of keywords and association analysis was performed using SAS Enterprise Miner 7.1, while the keyword network construction and network analysis was performed using NetMiner. In addition, the network constructed was schematized using NodeXL.

Before conducting the experiment, the target sample was refined in order to improve the quality of the analysis and shorten the analysis time. In this process, a lexicon on national pending issue and a temporary R&D lexicon were used to eliminate the terms that are irrelevant to national pending issue or R&D. The issue lexicon is constructed based on the topic analysis of 11,160 cases that are selected from a variety of

documents such as policy documents, news articles, discussions, and columns. And then, the issue lexicon and the R&D lexicon are combined. Next, the combined lexicon is used as a start list in the process of text parsing. Thus, it is possible to obtain a parsing result that comprises only issue keywords and R&D keywords. Further, based on this parsing result, the association rules between issue keywords and R&D keywords can be derived through association analysis.

A social network is constructed and analyzed based on the association rules derived above. In this study, a network is constructed based on the support value of the relevant keyword. An understanding of the relationship between issue keywords and R&D keywords that are highly associated with each other is helpful. Of the 10,000 rules derived in the first analysis, 1,000 rules with a support of more than 9% are extracted and a network is constructed based only on these rules. Each node is distinguished by the attribute name—"national pending issue," "R&D," and "national pending issue + R&D." Thus, it is possible to identify the meaning of each node. Figure 8 shows a part of the network derived above.



(Figure 8) A Part of Relevant Keyword Network using 1,000 Association Rules (Support Value of over 9%)

The circle part of Figure 8 indicates that the national pending issue keyword "copyright" is connected with R&D keywords such as "Digital," "Key," "Watermark," "Protection," and "Security". In this way, a set of R&D keywords corresponds to national pending issue keywords can be derived by analyzing the shortest distance and the maximum number of paths between the nodes.

4.3 Experiment: Constructing an Issue Network Based on Common R&D Keywords (Work in Progress)

This stage of the experiment involves two important processes. In the first process, the relationship between the issues and articles is derived through topic analysis of the news articles. In the second process, the relationship between the article and the R&D keywords is identified through frequency analysis of R&D terms obtained from each article. Next, two networks are constructed based on the results of these two processes. Lastly, a network between the issue and R&D keywords is constructed by merging these two networks. The relevant issues from the perspective of R&D keywords are identified through clustering analysis of the merged network.

The news articles and R&D lexicon are required to perform this experiment. Thus, we collected 13,652 articles published between June 2012 and July 2013 in the life/culture section of the South Korean news portal NAVER News. Thereafter, 1,000 issue topics are derived through topic analysis of these articles. In addition, using Daily Necessity as the main index, 10,012 cases from July 2012 to September 2013 are obtained from the patent information registered in South Korea. Based on the analysis of the collected patent information, an R&D lexicon that contains 3,040 terms is constructed. However, this lexicon contains not only R&D terms, but also general terms. The refinement of the lexicon through deletion of the general terms is still work in progress.

5. Conclusions

This study aimed to satisfy three objectives. The first objective was to overcome the limitations of the top-down and data-driven approaches in issue selection. The second objective was to address the discrepancies between the general issue keywords recognized by users and R&D keywords as well as package appropriate R&D information pertaining to specific issues. The third objective was to identify a set of related issues from an R&D perspective. Thus, we developed methodologies to address each objective.

Regarding the first objective of this study, the enhancement rate by the proposed integration methodology reveals to be about 42.8%. The enhancement rate seems to be different depending on the characteristics of each strategy (i.e. a group of issues). To address the second objective of this study, three

key analyses were conducted and a number of association rules between national pending issue keywords and R&D keywords were derived. The first analysis is the topic analysis of 11,160 cases obtained from market data such as policy documents, news articles, discussions, and columns. The other two analyses include association analysis of R&D keywords and social network analysis. Regarding the third objective of this study, the experiment is still work in progress. This experiment of issue clustering based on R&D keywords is expected to give tangible results in the future.

Lastly, three key directions are suggested for future research. First, since the entire analysis is based on topic analysis, it is essential to consider the semantic of each term. Second, since the analysis results may vary depending on the characteristics of the target documents, the analysis must be performed using various types and number of documents. Third, it is necessary to automatize the human intervention at each step of the analysis in order to improve the applicability of the proposed methodology. Finally, more precise validation and verification based on mathematical theory are required in order to apply the proposed methodology.

References

[1] R. Albright, “Taming Text with the SVD,” SAS Institute Inc., 2004.
www.sas.com/apps/whitepapers/whitepaper.jsp?code=SDM5

[2] J. Han, and M. Kamber, “Data Mining: Concepts and Techniques,” 3rd ed., Morgan Kaufmann Publishers: Massachusetts, 2011.
http://web.engr.illinois.edu/~hanj/bk3/

[3] R. J. Mooney, and R. Bunescu, “Mining Knowledge from Text Using Information Extraction,” ACM SIGKDD Explorations, Vol. 7, No. 1, pp. 3-10, 2005.
http://www.cs.utexas.edu/~ai-lab/pubs/text-kddexplore-05.pdf

[4] I. H. Witten, “Text Mining,” Practical Handbook of Internet Computing, edited by M. P. Singh, Chapman & Hall/ CRC Press, 2005.
http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf

[5] G. Salton, A. Wong, and C. S. Yang, ″A Vector Space Model for Automatic Indexing,” Communications of the ACM, Vol. 18, No. 11, pp. 613 - 620, 1975.
http://dx.doi.org/10.1145/361219.361220

[6] A. Stanvrianou, P. Andritsos, and N. Nicoloyannis, “Overview and Semantic Issues of Text Mining,” ACM SIGMOD Record, Vol. 36, No. 3, pp. 23-34, 2007.
http://dx.doi.org/10.1145/1324185.1324190

[7] E. Yu, J. Kim, C. Lee, and N. Kim, “Using Ontologies for Semantic Text Mining,” The Journal of Information Systems, Vol. 21, No. 3, pp. 137-161, 2012.
http://dx.doi.org/10.5859/KAIS.2012.21.3.137

[8] D. Jeong, M. Hwang, M. Cho, H. Jung, S. Yoon, K. Kim, and P. Kim, “Ontology and Text Mining-based Advanced Historical People Finding Service,” Journal of Internet Computing and Services, Vol. 13, No. 5, pp. 33-43, 2012.
http://dx.doi.org/10.7472/jksii.2012.13.5.33

[9] R. Agrawal, and R. Srikant, “Fast Algorithms for Mining Association Rules,” International Conference on Very Large Data Bases, Santiago, Chile, pp. 487-499, 1994.
http://dl.acm.org/citation.cfm?id=645920.672836&coll=DL&dl=ACM&CFID=656631652&CFTOKEN=27677818

[10] I. Cho, and N. Kim, “Recommending Core and Connecting Keywords of Research Area Using Social Network and Data Mining Techniques,” Journal of Intelligence and Information Systems, Vol. 17, No. 1, pp. 127-138, 2011.
http://www.dbpia.co.kr/Journal/ArticleDetail/1477011

[11] Y. Sohn, I. Kim, and N. Kim, “Automated Conceptual Data Modeling Using Association Rule Mining,” The Journal of information systems, Vol. 18, No. 4, pp. 59-86, 2009.
http://dx.doi.org/10.5859/KAIS.2009.18.4.059

[12] N. Kim, “Effect of Market Basket Size on the Accuracy of Association Rule Measures,” The journal of MIS research, Vol. 18, No. 2, pp. 95-114, 2008.
http://scholar.ndsl.kr/schDetail.do

[13] H. Ahn, I. Han, and N. Kim, “The Product Recommender System Combining Association Rules and Classification Models: The Case of G Internet Shopping Mall,” Information Systems Review, Vol. 8, No. 1, pp. 181-201, 2006.
http://dx.doi.org/10.13088/jiis.2013.19.2.039

[14] S. Yoon, “Churn Prediction Model for Department Store Customers Using Data Mining Technique,” Asia Marketing Journal, Vol. 6, No. 4, pp. 45-72, 2005.
http://academic.naver.com/view.nhn?doc_id=11465855

[15] Y. Lee, and K. Kim, “Product Recommender Systems using Multi-Model Ensemble Techniques,” Journal of Intelligence and Information Systems, Vol. 19, No. 2, pp. 39-54, 2013.
http://www.dbpia.co.kr/Article/3219909

[16] W. F. Wang, Y. L. Chung, M. H. Hus, and A. C. Keh, “A Personalized Recommender System for the Cosmetic Business,” Expert Systems with Applications, Vol. 26, No. 3, pp. 427-434, 2007.
http://dx.doi.org/10.1016/j.eswa.2003.10.001

[17] Y. Kim, “Social Network Analysis,” Bakyoungsa: Seoul, 2003.
http://book.naver.com/bookdb/book_detail.nhn?bid=128306

[18] S. Kauffman, “The Origins of Order,” Oxford University Press: New York, 1993.
https://global.oup.com/academic/product/the-origins-of-order-9780195079517

[19] K. Kwahk, “Social Network Analysis,” Chungram: Seoul, 2013.
http://book.naver.com/bookdb/book_detail.nhn?bid=7462254

[20] S. Park, and K. P. Kim, “A Closeness Analysis Algorithm for Workflow-supported Social Networks,” Journal of Internet Computing and Services, Vol. 14, No. 5, pp. 77-85, 2013.
http://www.dbpia.co.kr/Article/3282313

[21] K. Lee, H. Namgoong, E. Kim, K. Lee, and H. Kim, “Analysis of Multi-Dimensional Interaction among SNS Users,” Journal of Internet Computing and Services, Vol. 12, No. 2, pp. 113-122, 2011.
http://www.dbpia.co.kr/Article/1464198

[22] A. Jin, J. Lee, and J. Lee, “Measuring Method of String Similarity for POI Data Retrieval,” Journal of KIISE: Computing Practices and Letters, Vol. 19, No. 4, pp. 177-185, 2013.
http://www.dbpia.co.kr/Article/3140094

[23] B. You, and K. Choi, “A Study on the Construction of the National R&D Knowledge Information: Mainly Focused on the Research Planning and Management,” Journal of the Korean Society for Library and Information Science, Vol. 38, No. 1, pp. 281-301, 2004.
http://www.dbpia.co.kr/Article/348410

[24] S. Shin, Y. Yoon, M. Yang, J. Kim, and K. Shon, “A Data Cleansing Strategy for Improving Data Quality of National R&D Information - Case Study of NTIS,” The Korean Society Of Computer And Information, Vol. 16, No. 6, 2011.
http://dx.doi.org/10.9708/jksci.2011.16.6.119

[25] Y. Hyun, H. Han, H. Choi, J. Park, K. Lee, K. Kwahk, and N. Kim, “Methodology Using Text Analysis for Packaging R&D Information Services on Pending National Issues,” Journal of Information Technology Applications & Management, Vol. 20, No. 3, pp. 231-257, 2013.
http://www.dbpia.co.kr/Article/3257838

[26] L. Kwon, and J. Kim, “A Study on the Establishment of Reference Linking System for National R&D Information,” Journal of Korea Contents Association, Vol. 8, No. 1, pp. 195-202, 2008.
http://www.dbpia.co.kr/Article/761180

[27] M. Yang, Y. Yoon, S. Shin, J. Kim, and K. Shon, “A Development of Expert Search Agent System using National R&D Human Information Database for NTIS,” Journal of Internet Computing and Services, Vol. 11, No. 2, pp. 285-286, 2010.
http://www.dbpia.co.kr/Article/1390628

[28] J. Scott, Social Network Analysis: A Handbook, SAGE: California, 2000.
http://www.amazon.com/Social-Network-Analysis-John-Scott/dp/1446209040

◐ 저 자 소 개 ◑

현 윤 진 (Yoonjin Hyun)
2013 B.A. in Business IT, Kookmin University, Korea.
2013 ~ Present M.S. in Business IT, Kookmin University, Seoul, Korea.
Research Interests: Text Mining, Data Mining, Opinion Mining
E-mail : yoonjin0630@kookmin.ac.kr

윌 리 엄 (William Wong Xiu Shun)
2011 B.S. (Hons) in Computer Science, Universiti Sains Malaysia, Pulau Pinang, Malaysia.
2012 ~ Present M.S. in Business IT, Kookmin University, Seoul, Korea.
Research Interests: Text Mining, Data Mining, Opinion Mining
E-mail : williamwong@kookmin.ac.kr

김 남 규 (Namgyu Kim)
1998 B.S. in Computer Engineering, Seoul National University, Korea.
2000 M.S. in Management Engineering, KAIST, Korea.
2007 Ph.D. in Management Engineering, KAIST, Korea.
2007 ~ Present Professor, School of MIS, Kookmin University, Korea.
Research Interests: Text Mining, Data Mining, Data Modeling
E-mail : ngkim@kookmin.ac.kr