International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:10, 2014

Issue Reorganization Using the Measure of Relevance

William Wong Xiu Shun, Yoonjin Hyun, Mingyu Kim, Seongi Choi, Namgyu Kim

***Abstract*—**The need to extract R&D keywords from issues and use them to retrieve R&D information is increasing rapidly. However, it is difficult to identify related issues or distinguish them. Although the similarity between issues cannot be identified, with an R&D lexicon, issues that always share the same R&D keywords can be determined. In detail, the R&D keywords that are associated with a particular issue imply the key technology elements that are needed to solve a particular issue.

Furthermore, the relationship among issues that share the same R&D keywords can be shown in a more systematic way by clustering them according to keywords. Thus, sharing R&D results and reusing R&D technology can be facilitated. Indirectly, redundant investment in R&D can be reduced as the relevant R&D information can be shared among corresponding issues and the reusability of related R&D can be improved. Therefore, a methodology to cluster issues from the perspective of common R&D keywords is proposed to satisfy these demands.

***Keywords*—**Clustering, Social Network Analysis, Text Mining, Topic Analysis.

I. Introduction

THE volume of unstructured text data generated by various social media has increased so rapidly, it is now difficult to handle using existing data analysis methods. The volume of unstructured text data is particularly large and complex; however, this large volume of text data might contain useful information. Therefore, researchers have been motivated to develop big data analysis techniques and analyze this data in different areas such as politics, economics, and business [1], [2].

Furthermore, according to the needs of the analysis, text can be expressed in various forms such as words, phrases, hierarchies, or vectors. Therefore, text mining techniques are increasingly used. The text mining application that has received the most attention in academia and industry is topic analysis, usually used to extract the main topics from a large volume of text documents. Topic analysis is also referred to as issue analysis or trend analysis according to the type of target document. Measures such as TF-IDF (Term Frequency – Inverse Document Frequency) are used to derive the keywords that represent each document. In particular, topic analysis makes it possible to express a state corresponding to multiple topics in a single document. As a result, topic analysis is well-regarded compared to traditional document clustering that only allows a one-to-one relationship between topic and document.

William Wong Xiu Shun, Yoonjin Hyun, Mingyu Kim, Seongi Choi, and Namgyu Kim are with the Graduate School of Business IT, Kookmin University, Seoul, 136-702 Republic of Korea (e-mail: williamwong@kookmin.ac.kr, yoonjin0630@kookmin.ac.kr, minduz88@kookmin.ac.kr, csy0000@kookmin.ac.kr, ngkim@kookmin.ac.kr, respectively).

As the number of issues derived by topic analysis can be very large, issue clustering has become the main concern of research. The high-level concept of a new issue can be derived through clustering based on already-discovered issues. For example, the issue cluster “Diet” can be created by integrating the issues “Exercise” and “Obesity.” This issue clustering is mainly based on the co-occurrence frequency of issue keywords. In detail, if keywords representing the issues “Exercise” and “Obesity” appear simultaneously in various documents, they can be considered to be strongly correlated. However, an association between issues that have a low co-occurrence frequency cannot be recognized using traditional issue clustering methods, even if those issues are in fact strongly related. This limitation is further explained in Fig. 1.

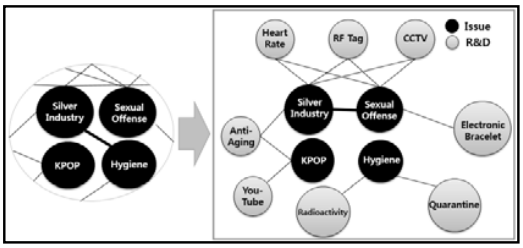

Fig. 1 Issue Clustering Network Diagram

Fig. 1 shows a network diagram of issue clustering from the perspective of related R&D keywords. If we examine the left hand side of the diagram, the issues “Silver Industry” and “Hygiene” are considered the strongly related keywords according to traditional issue clustering that considers only the co-occurrence of issues. The meaning of “Silver Industry” here indicates a business that focuses on products and services for senior citizens. However, according to the right-hand side of the diagram that considers both issues and the their related R&D keywords, “Silver Industry” and “Sexual offense” share three keywords related to technology, “Heart Rate,” “RF Tag,” and “CCTV.” From this association, it is clear that the R&D technology used in the “Silver Industry” field can also be applied in the field of “Sexual Offense.” Experts from these two field can then take advantage of this result and share R&D information among them. However, because the co-occurrence frequency of these two issues in the same document is not high, it is difficult to identify the correlation between them using traditional issue clustering. Thus, in order to overcome this limitation, a methodology to cluster issues from the perspective of related R&D keywords is proposed.

II. Related Work

A. Text Mining and Topic Analysis

Text is the most common form of information accessed on the Internet and smart devices [3]. Thus, many attempts have been made to analyze this large volume of text in order to extract meaningful information from different perspectives. Concern about topic analysis has increased recently, as it can identify the corresponding topic for a particular document by discovering the core keywords of a large number of documents. In addition, the widespread use of smart devices gives users the ability to express their opinions in text form through social media. Thus, various attempts have been actively carried out to establish business strategy through text analysis on social media data [4]. In addition, topic analysis can be applied to most areas that deal with text data, such as online shopping mall product review analysis, crime prediction, and repository establishment through text categorization [5]–[8].

Topic analysis is a comprehensive technique that had been utilized in many areas such as data mining, natural language processing, information retrieval, computational linguistics, and topic tracking. It plays an especially important role in natural language processing, as the target of the natural language processing, i.e., text, can be analyzed in various forms such as words, phrases, hierarchies, and vectors according to the purpose of the analysis [9]. Basically, the key concepts of topic analysis are the vector space model and TF-IDF metric. In detail, the smallest unit of analysis for each document is data written in text form, such as the text in a document title, summary, or body. Each document is represented using a vector space model; therefore, the subject and attributes of a particular document can be summarized according to the frequency of the terms used in each document. Analysis based on TF-IDF is more widely used than simple term frequency. Once the keywords are refined based on specific conditions, the co-occurrence patterns for each keyword are calculated based on the TF-IDF weights. However, because the number of terms used in a document is very large, in order to measure the similarity between documents, each document is stored after undergoing a dimensional reduction technique such as Singular Value Decomposition.

After the structuring of an unstructured text document is complete, the remaining topic analysis steps employ existing data mining methods. In particular, by clustering on document vectors, the technique of grouping similar documents is already being used in several areas. However, existing document clustering methods assume that each document only belongs to one cluster, and thus are too limited for actual topic analysis, as they cannot classify the document into various topics. Therefore, topic analysis, which allows many-to-many relationships between topics and documents, better reflects the actual state that corresponds to the multiple topics included in a complex document.

B. Social Network Analysis

Social network analysis is a quantitative analysis technique that identifies the characteristics of the connection structure of a group as well as the connection state of the objects within through visual representation [10]. Initially, it was used for the analysis of relationships among people. It is now also actively used for structural analysis in various fields such as genetic, transportation, and organization networks [11]–[13]. Recently, the range of applications is expanding as new knowledge is created through a combination of other techniques such as knowledge transfer factor analysis, keyword relationship analysis, and R&D information packaging [14]–[16].

In order to understand the link structure characteristics of a social network, various indicators such as density, centrality, and centralization have been discovered and widely used [17]. Density refers to the degree of connection between network nodes. High network density indicates that information exchange is active and information spread is fast. Centrality is an index that measures the extent to which a particular actor is connected to others in a network. It is subdivided into four main centralities: degree, closeness, betweenness, and eigenvector. In addition, centralization is an index that represents the degree to which the entire network is concentrated on a node; it is divided into degree centralization, closeness centralization, and betweenness centralization.

Tools such as NodeXL, UCINET, and NetMiner are frequently used for social network analysis [18]–[20]. These tools use a matrix to represent the data. A relationship between the rows and columns is expressed by 1; if there is no relationship, the value is 0. In addition, a quasi-network can be generated by deriving indirect relationships based on direct relationships among the objects [21]. For example, in a binary customer/product matrix where 1 represents a purchase and 0 represents a non-purchase, it is possible to construct a quasi-network between the customers by connecting the customers who purchased one or more of the same product. In this study, we built a two-mode network between news and R&D information as well as between news and issues. A two-mode network between R&D information and issues was constructed after merging the two networks such that further analysis is able to use the converted quasi-network between issues.

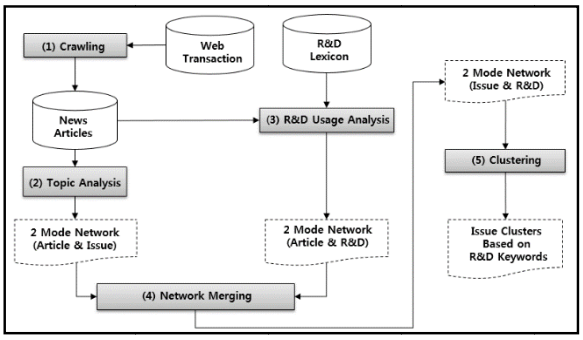

Fig. 2 Research Overview

III. Proposed Model and Research Scope

In this section, we further explain the methodology of clustering issues from the perspective of related R&D

keywords. Fig. 2 gives an overview of the research.

This research involves five important processes. Web transaction records of each internet user were used as input data. News articles related to research were first collected through (1) web crawling. Based on the collected news articles, the corresponding relationships between the issues and articles were derived through (2) topic analysis. Next, the corresponding relationships between an article and R&D keywords were identified through the (3) frequency analysis of the R&D terms in each article. Two networks were constructed based on the results of steps (2) and (3). In the (4) network merging stage, a two-mode network between the issues and R&D keywords was constructed by merging these two networks. Finally, the issues relevant to related R&D keywords were identified through a (5) clustering analysis of the merged network.

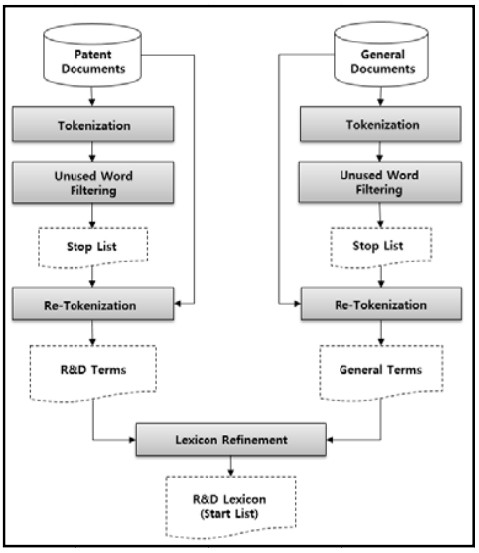

Fig. 3 R&D Lexicon Construction

Because the purpose of this study is to achieve issue clustering from the R&D perspective, it is essential to first construct an R&D lexicon based on R&D documents such as patent information. Fig. 3 shows the R&D lexicon construction process. First, the patent information is obtained and topics are then generated by parsing analysis (the tokenization stage). Each topic represents an R&D keyword. However, as the quality of the keywords is not sufficiently accurate, unused words are filtered. Therefore, word categories such as "URL," "time," "date," "number," or "meaningless word" are selected and included in the stop list. Using the stop list, the parsing analysis is generated again. A stop list was used here because it can be applied in other, similar experiments and reduces redundant calculation.

Thus, the results are a list of R&D terms. There are some general terms that always appear in patent information, so in order to remove these meaningless general terms from the R&D terms list, we generated a general terms list by the same process but using general documents as target data. Therefore, in the lexicon refinement stage, we remove all meaningless general terms that appear in the R&D terms list using the general terms list. The resulting refined R&D lexicon can be used as the initial list for R&D usage analysis.

IV. System Evaluation

A. Data Preprocessing

News articles and an R&D lexicon were required to perform this evaluation. 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 [22]. Thereafter, 100 issue topics were derived from topic analysis of these articles. In addition, using Daily Necessity as the main index, 10,012 cases from July 2012 to September 2013 were obtained from patents registered in WIPS, an online worldwide patent information service provider in South Korea [23]. Based on the analysis of the collected patent information, an R&D lexicon that contained 28,655 terms was constructed. However, this lexicon contained not only R&D terms, but also general terms. Therefore, the lexicon was refined by deleting general terms. After lexicon refinement, only 21,350 terms remained.

B. Topic Analysis

For this stage of our experiment, we used a commercial data mining tool, SAS Enterprise Miner 12.1 [24]. We performed topic analysis on the 13,652 news articles using its Text Miner module. The total number of topics was 100, and the analysis was limited to five keywords per topic. The results are shown in Fig. 4.

Topic ID	DocCutoff	TermCutoff	Topic Terms
7	0.455	0.021	village,forest,travel,scenery,course
11	0.435	0.019	symptom,treatment,patient,disease,immunity
16	0.343	0.018	skin,product,make-up,cosmetics,moisture
22	0.351	0.018	taste,food,restaurant,menu,ingredient
39	0.300	0.016	cancer,patient,breast cancer,surgery,inspection
51	0.286	0.015	stress,treatment,emotion,patient,confidence
63	0.211	0.014	alcohol,drinking,soju,concentration,BAC (blood alcohol concentration)
72	0.201	0.014	urine,disease,phlegm,symptom,nail
73	0.202	0.014	travel,product,tour,travel agency,hotel
79	0.201	0.014	movie,theater,audience,screening,cinema
87	0.130	0.014	male,female,patient,surgery,skin
89	0.140	0.013	healing,heal,kalium,natrium,mg
90	0.122	0.013	diet,coffee,beverage,weight,vitamin

Fig. 4 Topic Analysis Result (Part)

(a)

Doc.	Topic5	Topic6	Topic7	Topic8
1	0.053	0.197	0.242	0.116
2	0.159	0.125	0.175	0.370
3	0.313	0.547	0.465	1.492
4	0.092	0.190	0.189	0.349
5	0.067	0.145	0.107	0.265
6	1.200	0.144	0.342	0.346
7	0.109	0.445	0.125	0.416
8	0.192	0.810	0.194	0.578

(b)

Doc.	Topic5	Topic6	Topic7	Topic8
1	0	0	0	0
2	0	0	0	0
3	0	1	1	1
4	0	0	0	0
5	0	0	0	0
6	1	0	0	0
7	0	0	0	0
8	0	1	0	1

Fig. 5 Correspondence Matrix between Articles and Issue (Part)

When the many-to-many relationship between articles and issues were generated through topic analysis, the correspondence degree of each article for each issue was also identified. The initial correspondence degree is shown in Fig. 5 (a). Correspondence degrees above a certain threshold were set to 1; otherwise, they were set to 0. The converted

correspondence matrix between articles and issues is shown in Fig. 5 (b).

C. R&D Usage Analysis and Network Merging

In this section, we detail how to derive the two-mode network between articles and R&D keywords using the R&D usage analysis, i.e., the frequency analysis of R&D terms obtained from each article. In this study, matrix entries indicate the one-to-one relationship corresponding to a two-mode network. However, the two-mode network not only is complex, it is also difficult to explain when using only a small section. Therefore, the analysis process is explained using matrices with the same concept.

Fig. 6 describes the process of merging two different two-mode networks into one new two-mode network using matrices. Fig. 6 (a) shows the corresponding relationships between the articles and R&D keywords, while Fig. 6 (b) shows the corresponding relationships between the articles and issues. Using the number of articles that correspond to each R&D keyword and issue, the merged matrix in Fig. 6 (c) between R&D keywords and issues can be derived. This matrix explains the number of R&D keywords that appear to be related to each issue, also explaining the corresponding relationship between R&D keywords and issues. Using the generated final matrix, it is possible to derive a two-mode network using a threshold value. The correspondence degree above a certain threshold is converted to 1; otherwise, it is set to 0. Based on this procedure, a two-mode network between R&D keywords and issues can be constructed.

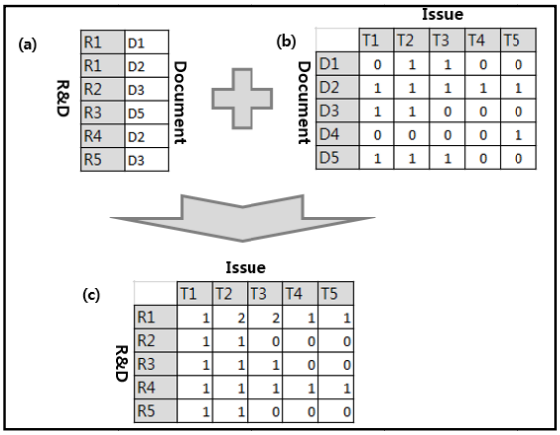

Fig. 6 Merging Example of Matrix (Network)

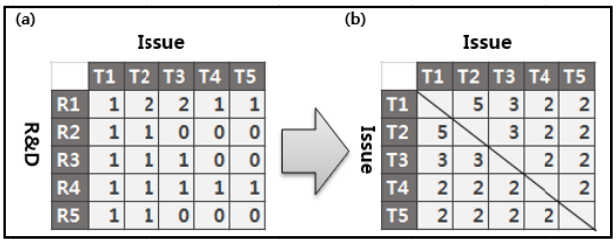

Fig. 7 Example Formation of a Quasi-Network between Issues

D. Quasi Network Formation

In this section, we generate a quasi-network between issues by converting the two-mode network between R&D keywords and issues. As in the previous section, the network conversion process shown in Fig. 7 is explained using matrices for clarity.

The matrix in Fig. 7 (a) is the same matrix as in Fig. 6 (c) and indicates the corresponding relationships between R&D keywords and issues. Among the five R&D keywords that correspond to issue T1, it can be seen that issue T4 corresponds to two R&D keywords (R1 and R4). In other words, two R&D keywords are shared by issues T1 and T4. Thus, using this method, it is possible to determine the number of R&D keywords that are shared by all issues. The results are summarized and shown in Fig. 7 (b). In this matrix, larger cell values indicate higher relevance between two topics. Thus, based on this final matrix, the quasi-network between issues was constructed and the final results are shown in Fig. 8.

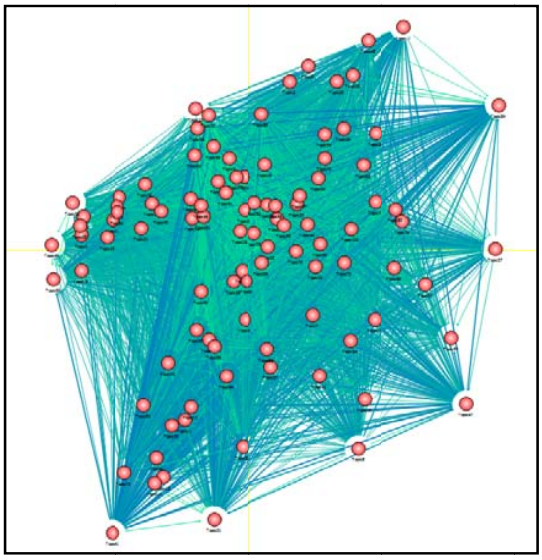

Fig. 8 Quasi-Network between the Issues

E. Issue Clustering

Based on the quasi-network between issues, clustering is performed. The links of the target network represent the R&D keywords that are often shared by two nodes (where a node represents an issue). Thus, based on the generated network, the clusters can be derived by grouping the issues that share a large number of common R&D keywords.

In this experiment, the clustering was performed using two different methods. For the first method, we clustered on the quasi-network of issues shown in Fig. 7 (b). Further, the clustering and visualization of the result was performed using NetMiner 4, a social network analysis and visualization tool. As shown in Fig. 9, the proximity between the issues was measured using the Euclidean distance matrix that corresponds to the network shown in Fig. 8.

		1	2	3	4	5	6
		Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
1	Topic1		35.8	26.6	34.2	36.4	37.0
2	Topic2	35.8		34.1	32.2	19.5	39.8
3	Topic3	26.6	34.1		34.8	33.3	36.6
4	Topic4	34.2	32.2	34.8		33.8	39.4
5	Topic5	36.4	19.5	33.3	33.8		41.1
6	Topic6	37.0	39.8	36.6	39.4	41.1	
7	Topic7	35.9	37.8	37.3	32.6	38.3	37.9
8	Topic8	38.9	40.6	39.1	38.9	40.3	36.2
9	Topic9	35.7	34.7	36.9	26.1	37.2	39.9
10	Topic10	34.4	33.1	35.2	30.3	35.5	39.9
11	Topic11	36.5	38.8	36.5	37.6	40.3	26.7
12	Topic12	37.4	40.8	35.5	41.2	41.0	21.9
13	Topic13	35.7	38.6	35.6	37.3	40.7	25.6
14	Topic14	38.4	41.4	37.0	41.2	42.1	23.7
15	Topic15	21.0	34.0	20.4	34.4	33.5	36.4
16	Topic16	35.5	36.5	36.4	28.1	37.1	33.7
17	Topic17	40.4	18.9	38.0	38.1	17.2	43.5
18	Topic18	33.3	33.4	33.5	25.9	35.1	39.1
19	Topic19	36.9	36.0	36.8	28.5	37.0	39.9
20	Topic20	37.0	38.6	38.0	37.7	39.1	37.7

Fig. 9 Euclidean Distancebetween Issues (Part)

Applying the algorithm provided in NetMiner, 20 clusters were derived using the Euclidean distance matrix shown in Fig. 9. Based on the 20 clusters generated, we discovered that most of the similar issues (topics) were grouped as a cluster. However, issues that did not appear to be directly related also appeared in the same cluster. Some examples are displayed in Fig. 10.

Topicid	Topic Name	Cluster
Topic7	village, forest, travel, scenery, course	6
Topic22	taste, food, restaurant, menu, ingredient	6
Topic65	festival, cherry blossoms, bicycle, event, spring flower	6
Topic73	travel, product, tour, travel agency, hotel	6
Topic79	movie, theater, audience, screening, cinema	6
Topic11	symptom, treatment, patient, disease, immunity	9
Topic13	pain, exercise, muscle, waist, joint	9
Topic14	vitamin, ingredient, health, groceries, vegetables	9
Topic39	cancer, patient, breast cancer, surgery, inspection	9
Topic51	stress, treatment, emotion, patient, confidence	9
Topic63	alcohol, drinking, Soju, concentration, BAC (blood alcohol concentration)	9
Topic72	urine, disease, phlegm, symptom, nail	9
Topic75	coffee, caffeine, beverage, ingestion, mg	9
Topic87	male, female, patient, surgery, skin	9
Topic94	constipation, prenancy, iron supplements, female, large intestine	9
Topic98	tooth, bacteria, gum, dentistry, cavity	9
Topic16	skin, product, make-up, cosmetics, moisture	10
Topic19	ingredient, salt, crushed, onion, garlic	10
Topic89	healing, heal, kalium, natrium, mg	10
Topic90	diet, coffee, beverage, weight, vitamin	10

Fig. 10 NetMiner Clustering Result (Part)

As shown in Fig. 10, many issues in cluster 6 seem to be related to each other, such as “travel,” “restaurant,” “festival,” and “hotel.” But topic 79, comprising issues “movie,” “theater,” and “cinema,” seems not to be related to the other topics in cluster 6. Further, the issues of cluster 10 such as “cosmetics,” “onion,” “healing,” and “diet” are unrelated to each other.

However, this phenomenon occurs because all these issues are grouped into a cluster based on common R&D keywords that they share. Therefore, in order to discover the relationship between the issues in a particular cluster, we derived the top 15 common R&D keywords that are shared by the issues in clusters 6, 9, and 10. The results are shown in Fig. 11.

Using cluster 6 as an example, the issues of topic 7, 22, 65, 73, and 79 such as “forest,” “restaurant,” “festival,” “hotel,” “cinema,” and “theater” can be group together by using the R&D keyword “fire extinguisher,” as it is very important to install fire extinguishers in the places listed above.

Furthermore, the management departments from these different fields can take advantage of this result to share R&D information. For example, the fire extinguishing system of the hotel can be reused and applied to a cinema or theater in order to upgrade the safety of the building and its people.

Common R&D Keywords of Cluster 6

imaginary space, mineral water, heat-resistant, immune globulin, microbiology, backlight, obesity, fire extinguisher, strain, visualization, embossing, insolation, natural light, indicator board, microbial substance

Common R&D Keywords of Cluster 9

polysaccharide, retort, microwave, asepsis, cellulose, dishwasher, acetyl, ethylene, medical appliances, chlorogenic acid, tapioca, calcium carbonate, tyramine, blood coagulation, plasma

Common R&D Keywords of Cluster 10

root-crop, laminariaceae, self-springing, salmonella, stain, slit, nutriments, lubrication, intravenous injection, extract, catheter, tense, phenolic acid, suspension, thrombosis

Fig. 11 Common R&D keywords of Cluster 6, 9, and 10

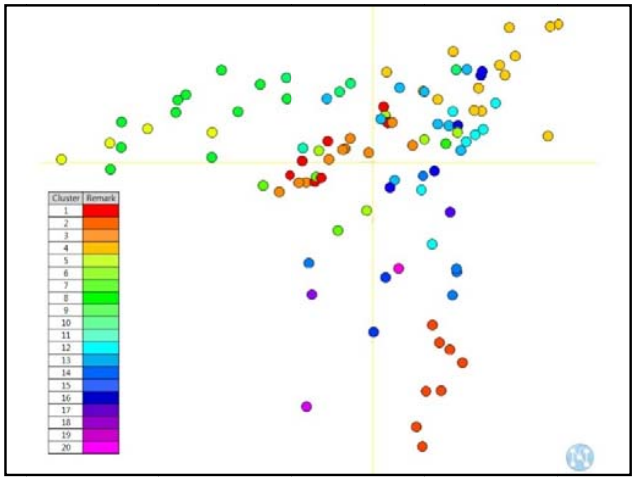

Fig. 12 Visualization Diagram of NetMiner Clustering Result

In addition, the visualization of the results in Fig. 10 was performed and the resulting visualization diagram is shown in Fig. 12. On the left bottom side of the diagram, the legend

indicates the 20 different colors corresponding to the clusters. In detail, issues of the same cluster are represented by the same color.

Next, a second clustering method was performed using SAS Enterprise Miner 12.1 based on the correspondence matrix between R&D keywords and issues. Before clustering, the rows and columns of the matrix shown in Fig. 7 (a) must be switched because SAS Enterprise Miner 12.1 can only perform clustering on matrix rows. A matrix with 100 issue rows and 2,380 R&D keyword columns, as shown in Fig. 13, was generated.

Topic_ID	Term44346	Term172283	Term172283	Term198215	Term198215
Topic1	0	2	2	0	0
Topic2	4	3	3	1	1
Topic3	2	2	2	0	0
Topic4	2	0	0	1	1
Topic5	2	3	3	1	1
Topic6	0	2	2	1	1
Topic7	0	0	0	2	2
Topic8	0	3	3	5	5
Topic9	1	2	2	1	1
Topic10	1	2	2	0	0
Topic11	1	2	2	1	1
Topic12	1	2	2	2	2

Fig. 13 SAS Clustering (Part)

In the SAS clustering, the number of clusters was set to 20. Because the number of clusters has a significant impact on the result, preliminary experiments to determine it were performed. However, a suitable number of clusters could not be determined this way. Thus, as an alternative, the number of clusters was randomly assigned to 20 based on the average number of issues included in a cluster. Example derived clusters are shown in Fig. 14.

Topicid	Topic Name	Cluster
Topic13	pain,exercise,muscle,waist,joint	3
Topic38	cigarette,non-smoking,smoking,smoker,non-smoking zone	3
Topic86	exercise,bicycle,walking,coffee,human	3
Topic87	male,female,patient,surgery,skin	3
Topic89	healing,heal,kalium,natrium,mg	3
Topic97	agency,seoul motor show,waist,specific agency,motor show	3
Topic4	color,style,fashion,item,one-piece dress	4
Topic16	skin,product,make-up,cosmetics,moisture	4
Topic11	symptom,treatment,patient,disease,immunity	12
Topic22	taste,food,restaurant,menu,ingredient	12
Topic98	tooth,bacteria,gum,dentistry,cavity	12

Fig. 14 SAS Clustering Result (Part)

For example, as shown in Fig. 14, the issues “exercise,” “smoking,” “surgery,” and “motor show” from cluster 3 seem to be unrelated. However, if these issues share the R&D keywords related to medical sensor devices such as “electronic tag,” then they are likely to be grouped together. This is also the case for cluster 12, where the issues “symptom,” “disease,” “food,” “bacteria,” and “dentistry” do not have a direct relationship among them, but by considering the related common R&D keywords that are shared among them, it is possible to find the relationship among these issues.

Thus, in order to discover the relationship between such issues, we derived their common R&D keywords. Using cluster 12 as an example, the top 10 common R&D keywords that were shared by topics 11, 22, and 98 were derived. As listed in Fig. 15, the common R&D keywords were “buffer action,” “carbon steel,” “synovial,” “electrolyte,” “root-crop,” “immune globulin,” and so on. Using the common R&D keywords, the relationship between these issues can be discovered. For example, “synovial” is the R&D keyword of topic 11, and usually refers to the synovial membrane or synovial joint. The synovial membrane is the soft tissue found between the joint capsule and joint cavity of synovial joints. However, “synovial” also appears as the R&D keyword of topic 98. To determine the relationship between the keyword “synovial” and topic 98, we searched the Internet using “synovial” and “dentistry” as search terms. We found that there is a rare case of synovial otorrhea of an iatrogenic nature in dentistry. Regarding the relationship between “synovial” and topic 22, we found that the synovial fluid was always connected with food, and “food” was one of the issues included in topic 22.

Topicid	Topic Name	Cluster
Topic11	symptom,treatment,patient,disease,immunity	12
Topic22	taste,food,restaurant,menu,ingredient	12
Topic98	tooth,bacteria,gum,dentistry,cavity	12

Common R&D Keywords
buffer action, carbon steel, synovial, electrolyte, root-crop,lubrication,catheter, immune globulin, larva,insolation

Fig. 15 Common R&D Keywords of Cluster 12

V. Conclusion

In this study, two clustering analysis tools were used to present interesting results and information. Given the two different tools used in this experiment, although the clustering results were different, there is potential to use the extracted issues for further research. In addition, sharing R&D results and reusing R&D technology can be facilitated. Indirectly, redundant investment in R&D can be reduced and the R&D information can be shared between those corresponding issues. Furthermore, the reusability of related R&D can be improved.

Three key directions are suggested for future research. First, because the entire analysis is based on topic analysis, it is essential to consider the semantics of each term. Second, because the analysis results may vary depending on the characteristics of the target documents, the analysis must be performed using various types and numbers of documents. Finally, it is necessary to automate steps of the analysis that are

currently manually performed in order to improve the applicability of the proposed methodology.

REFERENCES

[1] R. Albright, "Taming text with the SVD," SAS Institute Inc., Jan. 2004.
[2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann Publishers, 2011.
[3] I. H. Witten, "Text mining," PracticalHandbook of Internet Computing, CRC Press, 2004.
[4] I. Kim, "The Value of Big Data and Strategy," in *2012 Big Data Search Analysis Technology Insight*, 2012.
[5] J.Myung, D. Lee, and S. Lee., "A Korean Product Review Analysis System Using a Semi-Automatically Constructed Semantic Dictionary," Journal of KIISE: Software and Applications, vol.35, pp. 392-403, 2008.
[6] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.
[7] W. Fan, W. Wallace, S. Rich, and Z. Zhang, "Tapping the Power of Text Mining," *Communications of the ACM*, vol. 49, no. 9, pp. 76-82, 2006.
[8] F.Sebastisni, "Classification of Text," Automatic, *the Encyclopedia of Language and Linguistics*, 2nd ed., vol. 14, Elsevier Science Pub, 2006.
[9] A.Stanvrianou, P.Andritsos, and N.Nicoloyannis, "Overview and Semantic Issues of Text Mining," *ACM SIGMOD Record*, Vol. 36, pp. 23-24, 2007.
[10] Y. H. Kim, *Social Network Analysis*, Seoul, 2007.
[11] S.Kauffiman, *The Origins of Order*, Oxford University Press, 1993.
[12] S. Yoon, "A Study of Churn Prediction Model for Department Store Customers Using Data Mining Technique," *Asia Marketing Journal*, vol.6, no.4, pp. 45-72, 2005.
[13] C. Choi, "Research on Informal Organizational Network: Social Network Analysis," *Korea Society and Public Administration*, vol.17, no.1, pp. 1-23, 2006.
[14] M. Kang, and Y. S. Hau, "Multi-level Analysis of the Antecedents of Knowledge Transfer: Integration of Social Capital Theory and Social Network Theory," *Asia Pacific Journal of Information Systems*, vol.22, pp. 75-97, 2012.
[15] 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, pp. 127-138, 2011.
[16] Y. Hyun, H. Han, H. Choi, J. Park, K, Lee, K-Y. 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, pp. 231-257, 2013.
[17] K. Y.Kwak, *Social Network Analysis*, Cheongram, Seoul, 2014.
[18] NodeXL: Network Overview, Discovery and Exploration for Excel,Available at: http://nodexl.codeplex.com, Last accessed: 2nd August 2014.
[19] UCINET: UCINET Software, Available at: https://sites.google.com/site/ucinetsoftware/home, Last accessed: 2nd August 2014.
[20] NetMiner: NetMiner–Premier Software for Social Network Analysis,Available at: http://www.netminer.com, Last accessed: 2nd August 2014.
[21] S. Hong, *Social Network World and Big Data Applications*, Powerbook, Seoul, 2013, pp. 235-238.
[22] NAVER News,Available at: http://news.naver.com, Last accessed: 2nd August 2014.
[23] WIPSON,Available at: http://www.wipson.com, Last accessed: 2nd August 2014.
[24] SAS Software, Available at: http://www.sas.com, Last accessed: 2nd August 2014.