ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(23), DOI: 10.17485/ijst/2015/v8i23/79223, September 2015* ISSN (Online) : 0974-5645

A Study on the Construction of National R&D Data-based Customized Information Curation System

Tae-Hyun Kim, Myung-Seok Yang*, Nam-Gyu Kang and Kwang-Nam Choi

Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea;
msyang@kisti.re.kr

Abstract

Recently, there have been a lot of studies on contents curation for the purpose of overcoming the limitations of keyword-based general retrieval services. Contents curation is a part of information services provided by curators after classifying or processing related information in order to provide optimized data which meet each user's needs and demand instead of providing information to all users under the same method. This study proposes a national R&D data-based customized information curation system which sorts out target subjects among recent science and technology news and automatically extracts and curates related national R&D data with a goal of providing user-wanted information from the perspective of national R&D.

Keywords: Contents Curation, Information Curation System, Issue, National R&D Data

1. Introduction

In general retrieval services, query language-based search is conducted against the data indexed using a retrieval engine. Then, the results are classified by the type of data or provided in an integrated manner. For users to get wanted information under this kind of mechanism, they need to navigate all retrieval results or sort out them by editing or adding query language. Even though users find the information they want, they need to analyze or classify the data additionally to review the results in an integrated and comprehensive way.

Lately, there have been a lot of studies on contents curation for the purpose of overcoming the limitations of this kind of retrieval service. Contents curation refers to an activity to enhance the value of information by adding qualitative judgment on the information[1]. In other words, contents curation is a process of classifying or processing data with a goal of providing optimized data which meet users' need and demand, moving away from the conventional method in which data are provided to all users under the same method. Mobile app service, Kakao Topic provides customized news that reflect "my own perspective" and Pikicast provides fun and interesting contents that people love. This kind of mobile app service which sorts out the information and helps users get the information they want in a fast and accurate manner is a typical example of contents curation service.

This study attempted to search for a plan to curate and provide national R&D information relating to current science and technology news to provide customized information to users. This paper is structured as follows: In chapter 2, contents curation is more discussed, and national R&D information is explained. In chapter 3, studies on the extraction of issues and contents curation are reviewed. In chapter 4, how to construct a national R&D data-based customized information curation system is introduced. In chapter 5, the service that provides national R&D information relating to current issues using the system proposed in this study is briefly mentioned.

* *Author for correspondence*

In chapter 6, conclusion and future works are presented.

2. Background Knowledge

2.1 Contents Curation

In this big data era, a huge amount of information is produced every day. Under these circumstances, information finders are becoming increasingly difficult to get the information they want. This kind of information overload has even resulted in so-called 'the Hamlet Syndrome' in which people can't choose what they want because of too much information. As the acquisition and sharing of meaningful and valuable information become more important, contents curation has started to draw attention.

Contents curation service means a service that provides the contents users are related or interested in by collecting, sorting and editing various contents depending on their thought or perspective among a large amount of online contents. In the contents curation service, curators' role that analyzes these contents based on their own system and provides customized information by rearranging them by importance is critical. The pre-eminent publisher of magazines 'Reader's Digest' and news blog 'the Huffington Post' are the forerunners of curation contents. Main content curation services include Pinterest, Interest. me, Kakao Topic and so Pikicast. Content curation service has an advantage in that it can obtain the user's trust and sympathy by distributing the highly reliable contents based on the personal filtering[2].

2.2 National R&D Information

The National Science and Technology Information Service (NTIS) is the world's first national R&D information knowledge portal which provides national R&D program-related data such as programs, projects, human resources, equipment/facilities, and outcomes[3]. Since 2006, for the purpose of enhancing R&D efficiency from its planning to the utilization of outcome, the NTIS has provided a variety of services to researchers, policy makers and government officials by keeping national R&D information on database in linkage with 17 ministries and administrations (16 representative research management institutes) which have handled national R&D. To provide services after collecting national R&D information in a systematic and stable manner, the NTIS has defined the minimum national R&D information as 'Pan-bureau National R&D Information Standards', which is needed for the investigation of National R&D current status, comprehensive coordination and joint utilization from a pan-bureau perspective. Then, national R&D information has been linked and collected representative research management institutes[4].

3. Related Studies

3.1 Data Analysis and Curation

For the contents curation, it is needed to perform basic analysis such as extraction of keywords and calculation of weighted values and review curation-related studies which classify and process contents based on the results of the basic analysis and construct usable information.

Lee Sung-Jik, et al. adopted a keyword extraction method using an adjusted TF-IDF weighted model and vocabulary cross-comparison method in order to extract keywords from an optimum set of documents and classify them by the sector. This method can reveal large news articles on the Internet portals through a set of keywords. In the process, keywords equivalent to 'stop word' were removed through cross-comparison[5].

Lee Ki-Joon, et al. proposed a method which suggests a topic catalogue to users by clustering blogs based on topic similarities to make it possible to get access to wanted information in a fast and accurate manner in searching blogs. This method can classify and suggest related data even though the same keyword exists. Also, it ranks data by preference, popularity and reputation, allowing users to figure out the information they want more quickly[6].

Heo Jung, et al. suggested a method which constructs event templates by analyzing relevance information among keywords and provides changes in events and current information for the purpose of presenting an insight on social trend and changes in public opinion through analysis on contents such as social media, news and blog. This method is very helpful for users to make a decision because it provides relevance information among entities by time zone after overcoming the limitations of the keyword-based information supply system[7].

3.2 Implication

The said studies target to summarize or classify the results through the collection and analysis of the documents with a goal of extracting current topics from the news, blogs and tweets, or they focus on analyzing information itself by modeling and suggesting it in the form of a keyword network.

This study proposes a method to construct a curation system aimed to provide customized information based on the NTIS' national R&D data for the purpose of providing national R&D information associated with current social issues. In terms of extraction of topics relating to current social issues, a method that analyzes documents and extracts and clusters keyword, which has been found in conventional related studies has been used. However, this study attempted to provide national R&D information relating to current social issues in addition to the texts used in extracting these issues and trend information. For this, there has been an attempt to provide customized information through the automatic extraction and curation of information by utilizing metadata (e: classification information, keyword, etc.) in multiple aspects. To allow the curators to be able to curate the information which has been automatically extracted from topic extraction to the construction of related R&D information and trend information in a fast and accurate manner, a system that has specified curation process by stage has been designed.

4. National R&D Data-based Customized Information Curation System

4.1 Overview

The national R&D data-based customized information curation system proposed in this study has been designed as follows for the purpose of providing customized national R&D information relating to current social issues. It is structured as follows: 'document collection and management' stage in which texts such as news, policy briefing, and national policy report are collected and managed; 'document analysis' stage in which topic candidate information comprised of a set of keywords and classification information is automatically extracted through keyword clustering after analyzing collected texts and constructing a set of keywords by document; 'information curation' stage in which national R&D information and trend information relating to current social issues are automatically extracted after selecting target topics and constructing a set of filtered keywords through the curation of topic keywords. The details of the design by the stage are stated in the Sections 4.3 thru 4.5 below:

4.2 Curation-Targeted Information

The issue-related national R&D information refers to those closely associated with current national or social issues. If a curation process is used in providing this kind of information, it can provide a set of selected and processed compact information. Also, a new type of information can be added, or information can be processed and edited to a wanted format, enhancing users' satisfaction. The issue-related information defined for curation is shown in the table below:

By analyzing the collected original document from an external site, the issue is extracted. And then we use a set

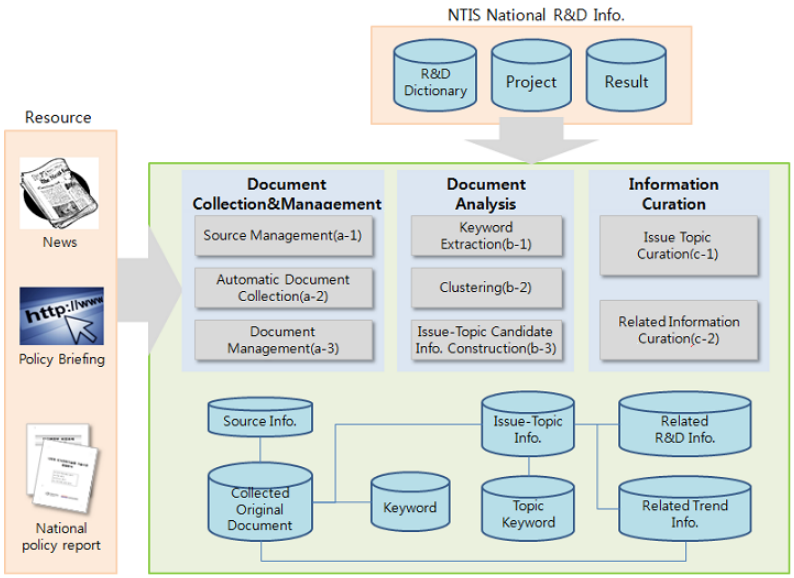

Figure 1. National R&D data-based customized information curation system.

Table 1. Curation-targeted information

Target Information	Description
List of issue-extracted texts	A list of texts associated with the issues among the texts collected from the news, policy briefing and national policy documents for the extraction of current national and social issues.
Set of issue keywords	A set of topic keywords among those extracted from the collected texts.
National R&D information	A list of issue-related national R&D projects and outcomes.
Set of R&D keywords	A set of issue keyword-related R&D keywords.
Trend information	A list of issue-related technology and policy trend information.

of issue keywords to express the meaning of the extracted issue. If these issues are expressed using a single keyword, however, it is hard to deliver the topic of the issue in a clear manner. Hence, a method presenting a set of keywords is applied. Also, unlike the common terms used in the news, technical words are often chosen in national R&D. Therefore, it is attempted to improve users' understanding of these R&D terms by additionally providing a set of related R&D keywords. The national R&D information, core information relating to current issues, is comprised of project information and outcome information that are available as fundamental data that provide yearly lists and various graphs. Furthermore, technology and policy trend information collected from the outside are additionally provided.

4.3 Document Collection and Management

To extract the issues closely related with national R&D information, this study investigated websites from which current national and social issues can be acquired, focusing on science and technology-related bureaus and news websites. The information in one of these websites was categorized into policy trend, technology trend, national issue or social issue and then the text issuance cycle and existence of RSS were examined. After applying the selection criteria stated in the table below, then, a total of 355 target documents and websites were confirmed.

This study extracted the topics of issues after automatically collecting texts from the policy trend and social issue-related websites every day. Because the information provided by the technology trend-oriented websites does not reflect social trends, however, it was used for the purpose of providing information on the related technology trends only, not as a target to extract the topics of issues. Since pending documents are usually made out and provided at a time at the establishment of a policy, they were excluded from the target of automatic collection. Instead, a data collection method was defined in a way to have the texts registered separately when they occur.

Based on survey results, information structure needed for document collection and management was defined. Then, source management, automatic document collection, and document management functions were designed. First, 'source management (a-1)' offers the registration and management of source information (URL, document, etc.) needed to collect the documents for the extraction of the topics of issues or policy and technology trend documents. Second, 'automatic document collection (a-2)' enables the automatic and periodic collection of documents using the registered source information. Lastly, 'document management (a-3)' offers the management of the automatically/manually registered documents by the date of registration.

4.4 Document Analysis

To extract issues, it is needed to derive keywords from the texts and extract candidates for issue topics through clustering. In 'keyword extraction (b-1)', keywords are extracted by analyzing morphemes against the collected texts, and the frequency of keywords by the collected text is calculated. Then, the keywords with low frequency and

Table 2. Target website selection criteria

Category	Selection Criteria
Up-to-dateness	The latest data only (collected for the last one year or less).
Accessibility	The paid websites and those not suitable for automatic collection in terms of RSS method excluded.
Objectivity	The websites with the articles of many subjective opinions and debate-oriented ones excluded.
Redundancy	The websites that provide redundant news under a different title and those providing the summarized forms of the articles which have been already provided by other websites excluded.

stop words are excluded from the keyword for analysis target. In 'clustering (b-2)', keyword clustering ('K-means' logic applied) is conducted using the statistical analysis programming language 'R' as shown in Figure 2[10]. If a group of highly related keywords is extracted through clustering, a list of texts including those whose keywords exceed the required level (more than 2 cases in the Figure below) is allocated to the group. Each keyword group extracted during the clustering becomes a candidate for issue topics.

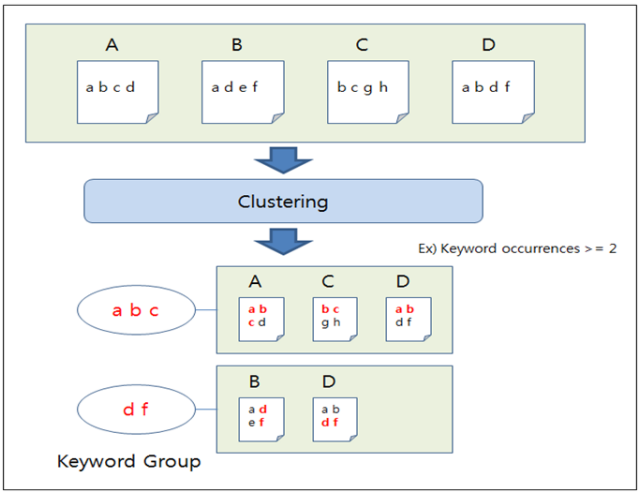

Figure 2. Example of keyword clustering.

In 'issue-topic candidate information construction (b-3)', keywords other than topic keywords are extracted from the related texts as shown in Figure 3 using the keyword group extracted during clustering and related texts and defined as 'general keywords'. After calculating the frequency of both general and topic keywords in the related texts, the information of the issue topic candidate is reinforced. This process allows a curator to get more useful information in selecting and processing service target information among the issue topic candidates.

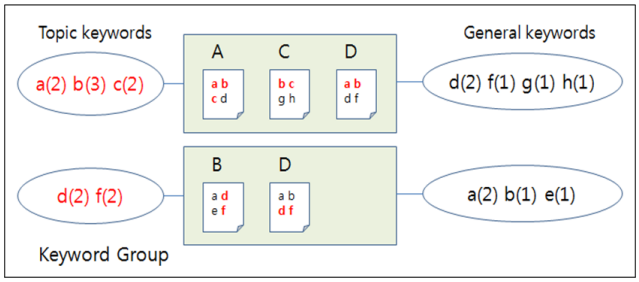

Figure 3. Construction of information of issue topic candidate

4.5 Information Curation

Information curation consists of issue topic curation which selects and filters target topic information using the information of issue topic candidates extracted during the document analysis and related information curation which automatically extracts related R&D information (national R&D project and outcome) and trend information and allows a curator to filter it. The 'issue topic curation (c-1)' offers a function to select target information that would be provided to users among the

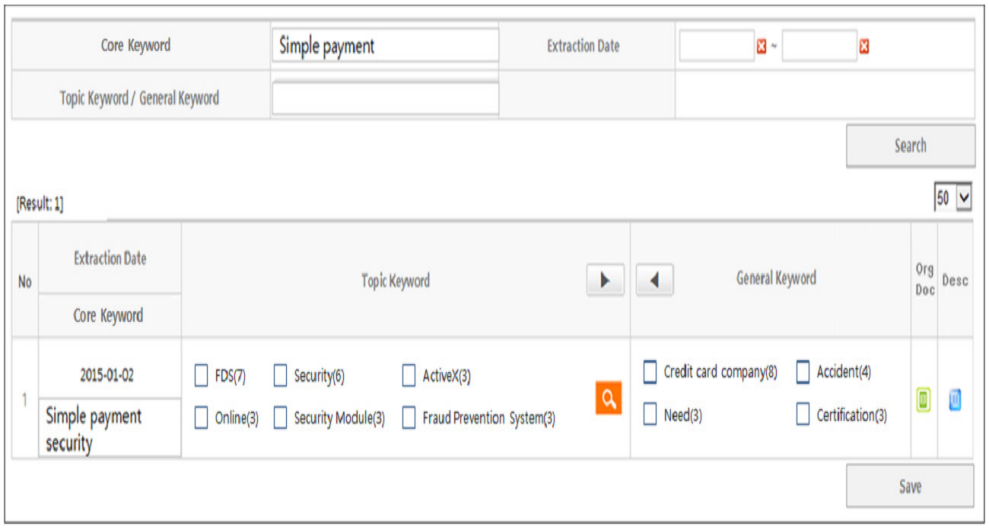

Figure 4. Management of issue topics.

information on issue topic candidates. Also, core keyword setting function and topic keyword adjusting function are provided in Figure 4. In the beginning, the keywords with the highest frequency among the topic keywords of issue topics are automatically set as the core keyword. A curator is allowed to change core keywords to more specific titles. Keywords are those which can enhance users' understanding after being presented along with core keywords when issue topics are provided. However, topic keywords automatically extracted through document analysis often have noise. Therefore, a curator should be able to change topic keywords into those who can express issue topics better by checking both general and topic keywords. For this, a function that can change keywords into general or topic keywords is provided. Furthermore, an issue topic curating function is optimized by providing a function which excludes the data in case unnecessary texts are included and the function that prepares an explanation by referring to an external website (ex: NAVER Encyclopedia and Wikipedia) to add a remark to the issue topic.

The 'related information curation (c-2)' supports the extraction and filtering of previous topics, R&D information and trend information relating to the target issue topics. It is divided into two stages: automatic extraction of related information; construction of the list of information to be provided by adding or deleting related information by allowing a curator to review the information.

In the stage of the automatic extraction of information, research field information (National Science and Technology standard classification, 6T classification) on each topic is extracted by matching selected topic keywords with R&D glossary (constructed using the keywords of national R&D projects, including R&D information such as each term's national science & technology standard classification and 6T classification, synonym and related word). Then, a list of related R&D keyword candidates is extracted from the R&D glossary, using the information relating to the synonym and related word of the topic keywords and R&D information. In this stage, a curator selects the R&D keywords that would be used in extracting R&D information after reviewing a list of related R&D keyword candidates. If there are the keywords to be added in addition to the automatically extracted R&D keywords, related R&D keyword curation processes such as direct retrieval and addition in the R&D glossary can be performed.

Once research fields, topic keywords, and related R&D keywords are determined, related previous topics, R&D information (project and outcome) and trend information are automatically extracted. Regarding the extraction of previous topics, in case of previous topics (issue topic information prior to the base date) which include some of the keywords based on those included in the issue topics, similarity scores were calculated (keyword weight x classification weight) by applying keyword and classification weights as stated in the Tables 3 and 4. Then, a list of previous topic candidates is suggested by listing the previous topics with higher scores on top.

Table 3. Application example of keyword weight values at issue topic comparison

		Comparison-targeted Topic		
		Topic Keywords	Related R&D Keywords	General Keywords
Issue Topics under Curating	Topic Keywords	5	5	3
	Related R&D Keywords	4	4	2
	General Keywords	2	2	1

Table 4. Application example of classification weight values at issue topic comparison

			Comparison-targeted Topic			Weight Value Applied
			National Science & Technology Standard Classification		6T	
			Main Category	Sub-Category		
Issue Topics under Curating	National Science & Technology standard Classification	Main Category	30			30
		Sub-Category	30	30		60
	6T				40	40
			30		40	70
			30	30	40	100

In the extraction of related R&D information, the following retrieval equation is constructed with the retrieval engine used by the NTIS using the research field, topic keywords, and related R&D keywords. Then, retrieval is carried out, and a list of project and outcome information candidates is extracted.

Retrieval equation: (Keyword #1 or ... or Keyword #N) AND (Category #1 or ... or Category #M)

In the extraction of related trend information, a list of technology and policy trend information is extracted by calculating similarity scores after applying keywords and classification weights just like the extraction of previous topics.

In the construction of service-targeted information, a curator constructs final service-targeted information by reviewing and filtering previous topics, R&D information, and trend information in sequence. In issue topic information, the same or similar topics are occasionally re-mentioned. If related R&D information and trend information on current issue topics are set just like the previous topics, therefore, it can construct a list of initial data on new issue topics more efficiently. For this, a curator is allowed to select wanted information among the data included in the previous topic and automatically take it over in the current issue topic information.

In the R&D information curation, it is permitted to select, add or delete the information suitable for the issue topic by checking the number of information and list by year as shown in Figure 6 below against project and outcome information which has been automatically extracted for the past five years for the convenience of information review. In the trend information curation, a curator can set a list of related trend information against technology and policy trend information for the past year.

In the information curation stage, a stage-by-stage information filtering function which enables the automatic extraction of issue topic-related information and allows a curator to check and review the extracted information with a goal of constructing a set of the most suitable information for each topic. Curators are given a clue to make a proper judgment, using research field information and keyword information in each sector in the NTIS. Also, they are allowed to carry out curation by stage, making it possible to promote national R&D data-based customized information curation in an organized and systematic manner.

5. Implementation

The issue-related national R&D information produced using the system developed in this study is provided through 'R&D information related to the issue' as shown in the figure below. The core keywords on top 10 latest issue topics were posted on the left top, allowing a user to select interested issues using menus. The core and topic keywords on the selected issues were located in the upper middle while the related R&D keywords and latest related issues were arranged on the right top, making it possible for a user to check the keywords at a glance. The R&D project and outcome information relating to

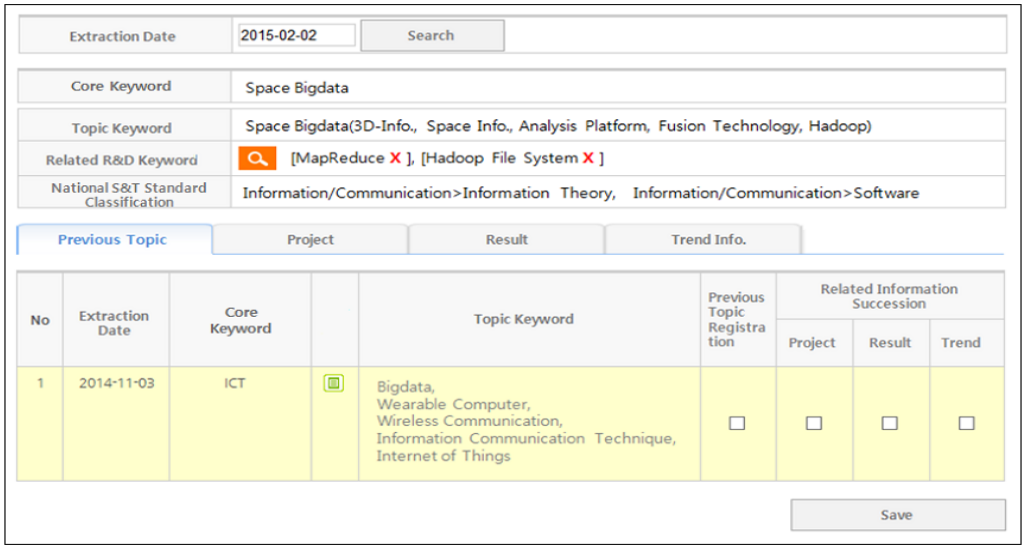

Figure 5. An example of the related previous topic setting.

Figure 6. A sample of 'R&D information related to the issue'.

the user-selected issue topic was posted in the middle of the screen, allowing a user to check it by the year or with a total list. If searched by the entire year, the number of projects by the year, research expenses, the number of participants and outcome table were provided to help users check the issue-related R&D information more quickly. Also, current status by the specific category, project performing agencies and list of senior researchers were provided by the project and outcome information for a more detailed investigation on research activities by the year. Furthermore, related trend information and other additional information (ex: explanation on issue topics, the list of texts, etc.) are provided.

6. Conclusion and Future Works

Keywords are extracted through analysis on current national and social issue documents collected from news and policy briefing websites and clustered. Then, issue topic candidates were acquired. After curators were allowed to provide functions to review these issue-topic candidates and filter topic keywords, the ground to provide curated national R&D information was established. The curation-designed information is provided to users in the following forms: keyword, the list of project and outcome and table of current conditions. As a result, general users, as well as professionals, can get the R&D information on their interests more easily. Moreover, it is likely that it would further increase in interest and investment in national R&D. To the future, to curating a similar analysis it needs to reflect national R&D information arising from a variety of fields, as well as issues, plans to structure as an independent component structure than the present system. To provide curated national R&D Information with more diverse perspectives, this study allows a user to understand national R&D information more easily and utilizes the information.

7. Acknowledgment

This research was supported by Maximize the Value of National Science and Technology by Strengthen

Sharing/Collaboration of National R&D Information funded by the Korea Institute of Science and Technology Information (KISTI).

8. References

1. You B-J. The future of information services. Etnews Contribution. 2015; 02(17):023.
2. Ji SJ. Contents curation service. Internet and Security Issue - NET Term. 2012; 12:26–30.
3. National Science & Technology Information Service (NTIS). Available from: http://www.ntis.go.kr
4. Yang M-S, Choi K-M, Jung O-N, Kim J. Some considerations on National Science & Technology Information Service (NTIS). Korea Technology Innovation Society 2013 Spring Conference; 2013. p. 294–304.
5. Lee S, Kim H-J. Keyword extraction from news corpus using modified TF-IDF. The Journal of Society for E-Business Studies. 2009; 14(4):59–73.
6. Lee KJ, Lee MJ, Kim WJ. Study for blog clustering method based on similarity of titles. Journal of Intelligence and Information Systems. 2009; 15(2):61–74.
7. Heo J, Ryu P-M, Choi YJ, Kim HK, Ock CY. An issue event search system based on big data for decision supporting: Social wisdom. Journal of KIISE. Software and Application. 2013; 40(7):381–94.
8. Kang B, Song M, Jho W. A study on opinion mining of newspaper texts based on topic modeling. Journal of the Korean Society for Library and Information Science. 2013; 47(4):315–34.
9. Jin SA, Heo GE, Jeong YK, Song M. Topic-network based topic shift detection on twitter. Journal of the Korean Society for Information Management. 2013; 30(1):285–302.
10. The R Project for Statistical Computing. Available from: https://www.r–project.org/
11. Yang M-S, Kang N-K, Kim T-H, Joo W-K, Park M-W, Choi K-N. An analysis of National R&D Collaborators Network based on the NTIS data. International Journal of Software Engineering and Its Applications. 2014; 8(11):11–24.
12. Kwak C-U, Yoon H-G, Park S-B. Query expansion based on word sense community. Journal of KIISE. 2014; 41(12):1058–65.
13. Ko EB, Lee JW. Sentence similarity measurement method using a set-based POI data Search. KIISE Transactions on Computing Practices. 2014 Dec; 20(12):711–6.
14. Choi H-S, Kim J-S. Design of a NTIS information integration model based on a standard integration platform. Journal of KIISE. 2012; 18(6):484–8.
15. Hong J, Choi H, Han H, Kim J, Yu E, Lim S, Kim N. A data analysis-based hybrid methodology for selecting pending national issue keywords. Entrue Journal of Information Technology. Special Issue. 2014; 13(1):97–111.
16. Hyun Y, Han H, Choi H, Park J, Lee K, Kwahk K-Y, Kim N. Methodology using text analysis for packaging R&D information services on pending national issues. Journal of Information Technology Applications and Management. 2013; 20(3):231–57.