An Evaluation of Applying Knowledge Base to Academic Information Service

Seok-Hyoung Lee*, Hwan-Min Kim**, Ho-Seop Choe***

ARTICLE INFO

Article history:
Received 28 March 2013
Revised 2 April 2013
Accepted 10 June 2013

Keywords:
Knowledge Base,
Usefulness Evaluation,
Academic Information Service,
URI Management,
Meaning Discrimination

ABSTRACT

Through a series of precise text handling processes, including automatic extraction of information from documents with knowledge from various fields, recognition of entity names, detection of core topics, analysis of the relations between the extracted information and topics, and automatic inference of new knowledge, the most efficient knowledge base of the relevant field is created, and plans to apply these to the information knowledge management and service are the core requirements necessary for intellectualization of information.

In this paper, the knowledge base, which is a necessary core resource and comprehensive technology for intellectualization of science and technology information, is described and the usability of academic information services using it is evaluated. The knowledge base proposed in this article is an amalgamation of information expression and knowledge storage, composed of identifying code systems from terms to documents, by integrating terminologies, word intelligent networks, topic networks, classification systems, and authority data.

1. Introduction

The web (World Wide Web: WWW) devised by Tim-Berners-Lee has become the footing for finding and utilizing wanted information online. The development of computers and networks, and the development of various standards including HTML, has hugely contributed to the sharing and exploring of a vast amount of information through the internet. But due to the exponential generation of information, it has become harder and harder to find any wanted information, and more effort must be put in for obtaining information.

* Senior Researcher, Department of Overseas Information, Korea Institute of Science and Technology Information; Adjunct Professor, Department of Library and Information Science, Konkuk University (skyi@kisti.re.kr)
** Senior Researcher, Department of Overseas Information, Korea Institute of Science and Technology Information (mrkim@kisti.re.kr)
*** Curatorial Official, Department of New Museum Planning & Construction, National Museum of Contemporary Art (hschoe@kisti.re.kr)
International Journal of Knowledge Content Development & Technology, 3(1): 81-95, 2013.
[http://dx.doi.org/10.5865/IJKCT.2013.3.1.081]

To find information using the keyword based search, which is the conventional information search method presently provided by portal sites such as Google, Naver, and Daum, the keyword matching documents have to be filtered from countless web documents and ranked using certain standards, and the user must go through each of the countless results listed in order. In particular, the science and technology field information services provide documents with huge amounts of terminologies, which renders it difficult for the users to obtain search results reflecting their intentions using primitive queries, and furthermore, even if the search results were obtained, the general users have to reconstitute their query to find the wanted document, or find documents through links in the related web pages, consuming both time and effort.

To resolve these problems, various studies on the intellectualization of information are being performed in the information science field. By intellectualizing information, meaning is granted to the information formed by simple strings of text, and by setting relationships, countless data can be expressed as knowledge information. But for the intellectualization of information in the information technology aspect, a considerable amount of effort is necessary, and technologies such as semantic web-related technology, natural language processing technology, next generation information searching technology, high precision information analysis technology, and personalization technology are required (Choe, 2004).

Based on these technologies, multi-dimensional approaches are being made to the methodologies of providing meaningful information to individuals or groups of people by systemic storage and management of information. From the knowledge information processing dimension, from the systematization of information, intellectualization of information, and automation of knowledge provision viewpoint, automatic extraction of information from documents with knowledge from various fields (information extraction), recognition of entity names [name of person·location·organization] (name-entity recognition), detection of core topics (topic detection), analysis of the relations between the extracted information and topics (relation analysis), and automatic inference of new knowledge (inference) consist the series of precise text handling processes, to try and create the most efficient knowledge bases for each field (Park, 2000; Seung, 2000; Rowley, 2008). Not only do the created knowledge bases analyze and provide meaningful and accurate knowledge information to the information users, they also provide opportunities for the information manager to systematically store and manage accurate information.

In this article, the authors have proposed the course of direction for science and technology academic information services utilizing the knowledge base, which is a necessary core resource and comprehensive technology for intellectualization of information, and evaluated its usability.

The composition of this article is as follows: Section 2 discusses the overall concept of the knowledge base, and in Section 3 the components of the knowledge base are described. In Section 4, the direction of information services utilizing the knowledge base is sought, and in Section 5, its usefulness is evaluated and the conclusions are made.

2. Related Work

2.1. Information Retrieval & Service

From the viewpoint of the user, information search technology acts as a medium to aid in providing meaningful knowledge information through various information searches. However, from the viewpoint of the developer, it is more focused on the descriptive processing of knowledge information, such as expression, storage, organization, and access of knowledge information, including handling of queries, large capacity information processing, and indexing technologies. Recently, research has been done on the methods of managing, storing, and servicing knowledge information, both semantically and in practice.

The recent progress in the research and development of information search and service technologies has been done actively in the area of semantic-based information search, including natural language processing technology, the CLARIT (Evans, 1996) or IRENA (Arampatzis, 1997) system using phrase-based indexing through syntax analysis of noun phrases in the document text, systems using indexing methods based on phrase structure analysis called TSAs (Tree Structure Analytics), and natural language processing technology, including sentence-based searching and question-and-answer type searching methods of the web through the establishment of necessary knowledge bases.

But semantic-based information searching is a time and effort consuming technology that requires the processing of vast amounts of stored information, establishment of a large capacity knowledge base, and precise language processing technologies. In this article, the authors propose academic information service functions based on knowledge bases allowing semantic information searches for users while maintaining the current knowledge information data structure.

2.2. Knowledge Base

Generally, the knowledge base is a component of expert systems. It is a database containing technical knowledge accumulated by the artificial intelligence agent through intellectual activity and experience related to the field to be used, and the facts and rules required to resolve problems (Brachman, 2004). Knowledge modularization and structuralization are done through the creation of technical knowledge and expression rules, and, as a result, knowledge addition and modification are made possible, enabling new knowledge to be expressed in the knowledge base due to changes in the situation.

The knowledge bases are applied to semantic analysis for language processing or at the inferencing stage, and these may be summarized in to lexical databases, formed from hierarchical structures of lexicons, such as the thesaurus or the ontology, and the fact databases or the real-world knowledge bases. The former is similar to the thesaurus, or the controlled vocabulary of the library and information sciences, and the ontology, or the semantic network of the computer sciences As to the latter, the CYC of Cycorp, which has expressed the vast human basic or “common sense” knowledge into a logical system, is a representative example of a real-world knowledge base.

The knowledge base of this article is based on the former knowledge base with the lexical database structure. The individual items of the knowledge base, including science and technology terminologies,

word intelligent networks, topic networks, classification systems, and authority data are organized to have structures mutually related between concepts (or entities), and these knowledge bases themselves are configured to cross-reference between individual entities based on identifying codes (Choe, 2004).

The structures of these science and technology knowledge bases are similar to those of the thesaurus, controlled vocabulary, or semantic network. In particular, out of the components of the knowledge base, the word intelligent network not only has the basic semantic relations and basic concept relations, it also has a conceptually expanded structure system according to these relations (Im, 2005).

3. Science and Technology Knowledge Base

3.1. Concept

Generally the knowledge base is a knowledge storage system connecting the vocabulary and documents of the field of science and technology, including terminologies, word intelligent networks, topic networks, classification systems, and author identifiers.

The inherent identifiable entities, terminologies, keywords, and classifications included in the science and technology information are connected closely to the knowledge base components including science and technology word intelligent networks, topic networks, classification systems, and author identifiers, and are categorized and identified according to its related field and stored in the knowledge base. The knowledge bases composed in this way are mutually interfaced and expressed in a knowledge expression form understandable to both computers and humans, and utilized not only for science and technology information management and service, but also for interfacing with other information resources and information analysis.

As previously stated, not only the science and technology academic documents, but also semantically-identified entities, including author names, institution names, journal titles, terminologies, research topics, and categorical information, are assigned proper IDs through URI managers, as shown in Figure 1, and the entities may be interfaced semantically using the assigned IDs. Utilizing these interfaced link structures, the user can obtain the only related information, beginning information searches from any of the search access points.

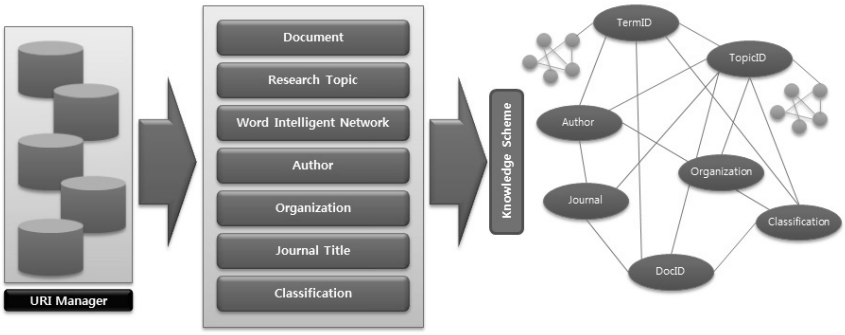

Fig. 1. Identifying code interface of the knowledge base

3.2. Components

In this passage the authors describe the components of the knowledge base, such as terminology, word intelligent network, topic network, classification systems, and author identifiers.

3.2.1. Terminology

Terminologies are terms used frequently in a certain specialized field, and, out of the vocabulary used in the relative field, indicate terms frequently used by experts. By using terminologies with inherent concept and technology, anyone can express and deliver knowledge freely, and it may be possible to perform expert communications, and to accumulate and understand knowledge freely using various communication methods.

The terminology, which is the base resource of the knowledge base, is a glossary consisting of words searched and collected from terminology dictionaries of the science and technology fields, science and technology academic information, and other web information, and established for each detailed field. These terminologies may be utilized not only for the knowledge base, but also for the thesaurus and controlled vocabulary, and may be utilized for detection and discovery of technologies through semantic analysis of documents and life cycle analysis of terms.

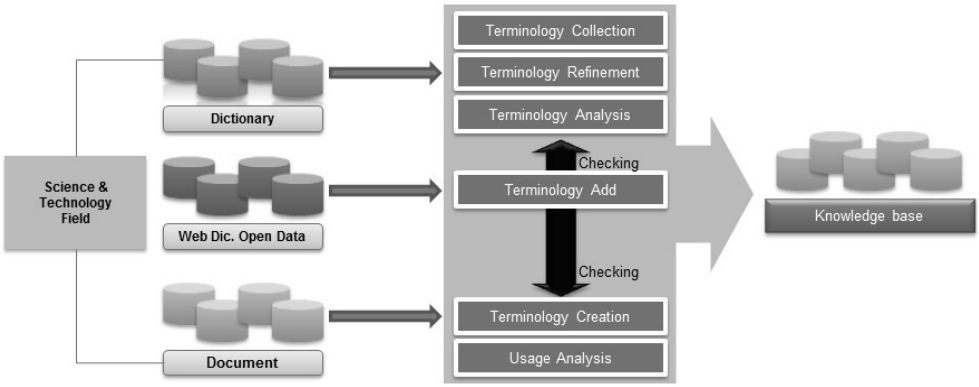

Fig. 2. Construction of terminology

3.2.2. Word Intelligent Network

The word intelligent network is a word database and network formed by comprehending the cognitive systems and conceptual relations of the Korean language based on the common and individual attributes, and forming these into a semantic and conceptual network of words (Choe, 2006). The word intelligent network includes the semantic relations (hierarchical relations, synonymous relations, whole and partial relations, etc.) and conceptual relations (attributes of semantic relations) between words, centered on the semantic relations of the hierarchy structure of general words, based on the terminology of the science and technology field.

For the word 'radiation therapy', which is a medical term, the 'treatment → method' may be configured as a general hyperonym, but utilizing the characteristics of the treatments in the medical

field and related document information, a detailed hierarchical structure was formed. Therefore, various types of 'physiotherapy' are present as 'phototherapy, electrotherapy, crenotherapy, climatotherapy, thermotherapy, massotherapy', and various types of 'phototherapy' are present as 'laser therapy, radiation therapy, artificial sunray therapy, ultraviolet therapy, infrared therapy', and using this medical field information, the hierarchical relations of the word 'radiation therapy' may be accessed in a Top-Down method and a Bottom-Up method, considering the unique characteristics of the expert field, and established as in Figure 3.

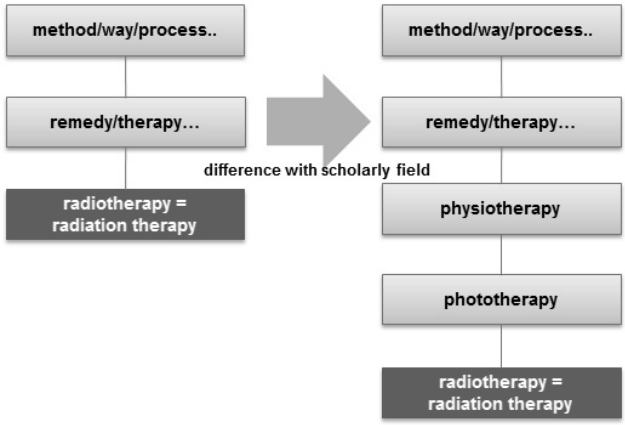

Fig. 3. The hierarchical structure of the word intelligent network

3.2.3. Topic Network

The science and technology topic network analyzes the keyword information of the academic information to implement a network of relations between topics. As shown in Figure 4, various keywords exist to represent the topic of scholar information, and in the topic network, relations between various keywords are digitized and expressed in network form.

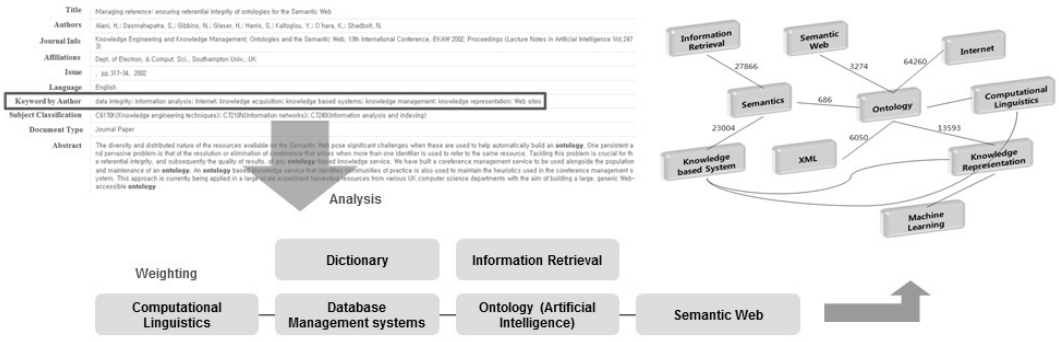

Fig. 4. Composition concept of the science and technology topic network

The relations between keywords may be expressed as the frequency of emergence of each keyword.

For example, if the terms “semantic web” and “ontology” are stated as keywords in 15 articles at the same time, a relation index of 15 is created between the two keywords. As higher relation indexes may stand for higher relations between two keywords, it may have value as a base resource for the topic network to provide more meaningful information to the users.

3.2.4. Authority Data

The authority data are records of source information of science and technology information, including individual names (or names of groups), signature, and topic, by finding all possible expressions to maintain consistency and coherence, confirming access points following certain rules, and configuring mutual references. Authority data is generally related to the entity name recognition and identification technology in the fields of computer science, but the essential difference is that the entity name recognition and identification is a technology to mostly sort the names, locations, and organizations based on certain machine learning. The authority data not only identifies proper entities, but configures allomorphs and source information semantically.

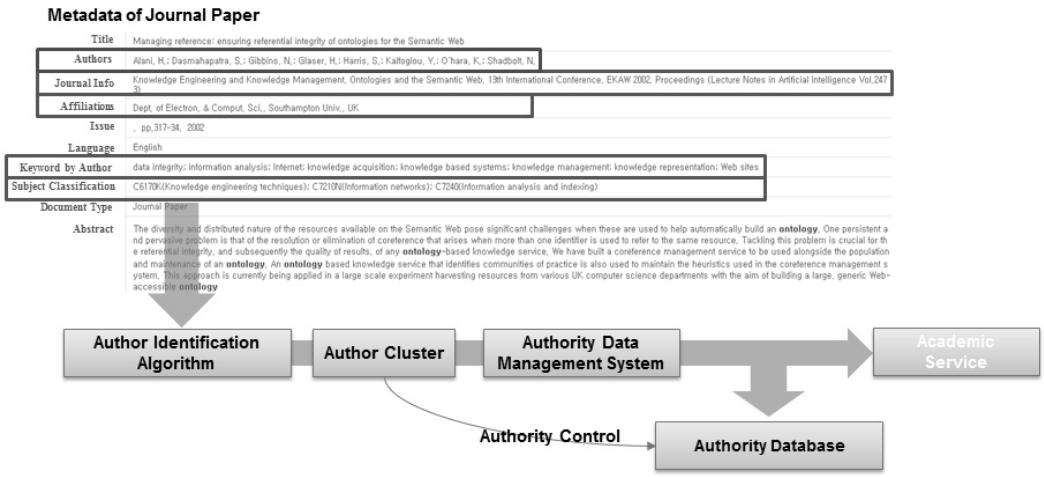

Fig. 5. Composition concept of the topic network

The authority data can be utilized for every step in a series of information searches, including query input, search performance, and combination of search results, and it is essential for providing accurate information. In particular, due to the exponentially increasing amount of information in academic information searches, the need for improving access control and providing information at the right time to the end user has grown more imperative, and the utilization of authority data has become very important to provide heteronyms and homographs, which are generated from different recognition of the same meaning between people, differentiating the meaning for the users.

3.2.5. Classification

All information is allocated systematically to an appropriate location according to a given principle, concept, format, or topic, and a classification system is then required for this purpose. The classification

system is a tool for conceptualizing specific information, an instrument for representing and describing knowledge and enhancing access convenience, and performs important functions for systematic information management and services. A classification system is thus applied and used with most of the information, and science and technology information also has a plurality of science and technical classifications.

In this study, we consider a mapping system for classification systems for connecting a plurality of science and technical classification. This system can be applied to relation types of classification classes with a multi-classification system connection structure on the basis of related studies about existing classification system connection structure, to support mapping and management of classification systems with a plurality of features.

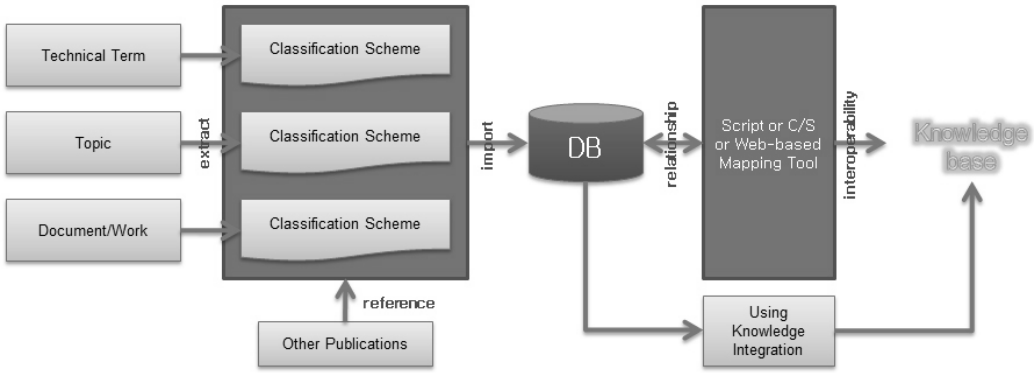

Fig. 6. Construction of Classification mapping

3.3. Management

The pre-existing management for academic information was performed for the purpose of simply collecting and accumulating data, to service information. For this reason, academic information management in the form of creating indexing structures for academic document titles, abstracts, keywords, and author name values, and referencing authority data and topic categories were common. But knowledge information management based on knowledge bases should support the user's process of searching and selecting information efficiently, and should be able to provide accurate and meaningful information to the user from any information search location (Zins, 2007). To achieve this, terminologies, word intelligent networks, topic networks, and authority data should all be interfaced with major metadata values of the target academic documents to be searched, and these knowledge bases should be established under a systemic management system.

As shown in Figure 7, the title, author, society, journal, keyword, topic category, and abstract information shown in the science and technology academic information are the search starting points for the user. For these items to be identified and connected semantically, this system should be configured and managed by knowledge bases.

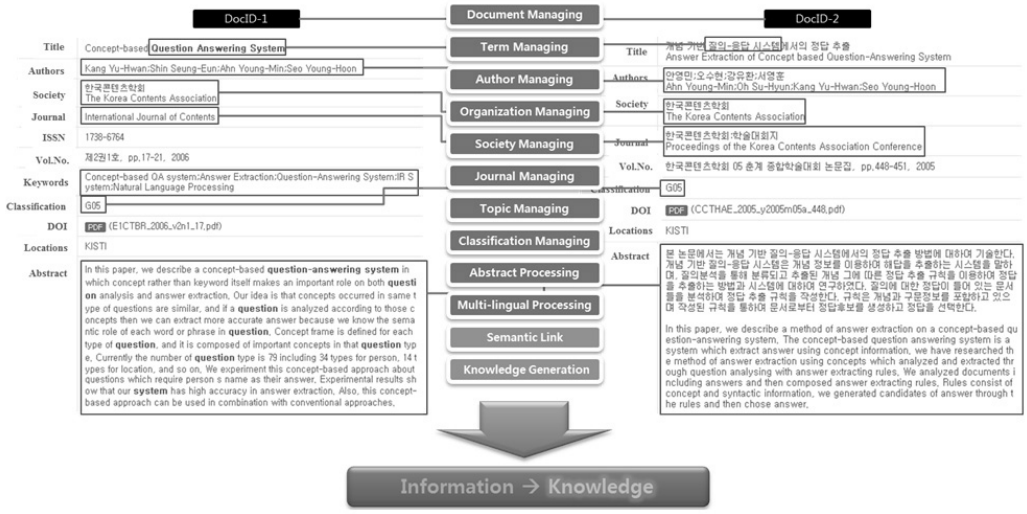

Fig. 7. Academic information management based on identification of knowledge base components

In Figure 7, management of the author names, institution names, and society names is done through the science and technology authority data system, and the terms in the titles and abstracts are interfaced with science and technology terminologies. The keywords are reconfigured to provide related information to the user through the science and technology topic networks, and the classification system is linked and managed with the science and technology classification system interface structures. Thus, the metadata items shown in one science and technology academic document have individual inherent meaning, and if these are linked to the components of the knowledge base 1:1, it should be possible to provide a better information service environment for the users.

To configure a management environment as written above, a knowledge base integrated management system is necessary. As shown in Figure 7, the academic information, terminologies, word intelligent networks, topic networks, authority data, and classification systems have to be managed based on identifiable codes (URI Manager) for the science and technology integrated management system. Considering the interface with the other heterogeneous systems, the various types of academic information and language resources should be managed in a common form (Metadata Registry). Also, thorough this management system, it should support terminology extraction, semantic network configuration of language resources, and interface functions (Relation Manager) with the academic information from the language resources emerging from academic information. The knowledge base should have a knowledge representation structure providing semantic interface systems between these knowledge base components. This information may be provided as personalized knowledge information fit for each individual's characteristics, and in this process, if individual knowledge bases are created and utilized that reflect the real-world knowledge by analysis of search patterns of experts, analysis of the tendencies of experts, and analysis of related topics, this may be applied to information services including real-time personalized information recommendation, and sharing and utilizing user information with similar tendencies.

4. Functional components of the academic information service

The most distinguished feature of knowledge base information service is that it can provide accurate and meaningful information to the user.

In this study, the authors have proposed two methods of service techniques that may apply the knowledge base from the viewpoint of the academic information query input stage and information search stage.

4.1. Query suggest function

The query suggest function is designed to rapidly input the name or address in the web browser or other software in case the same words are entered repeatedly. The relevant text is stored in the computer and compared to the newly entered text, and a list of representative terms is shown in the instant a word is input, and the relevant contents are selected and searched. In the case of information searches, it is a very convenient function to find the wanted queries by the user, and recently most search sites provide this function with the application of Ajax technologies.

At ordinary search sites, the implemented query suggest functions were implemented using the query history list, but the search suggest functions with knowledge bases applied utilizes the word intelligent networks, topic networks, and authority data to propose accurate queries at the beginning of a search.

The primary search suggest functions are provided using the terminology information, and using the hierarchical structures of the word intelligent network, secondary search suggest functions are provided for the subordinate queries.

The following figures are examples of the query suggest function with word intelligent network (Figure 8), topic network (Figure 9), and authority data (Figure 10) applied.

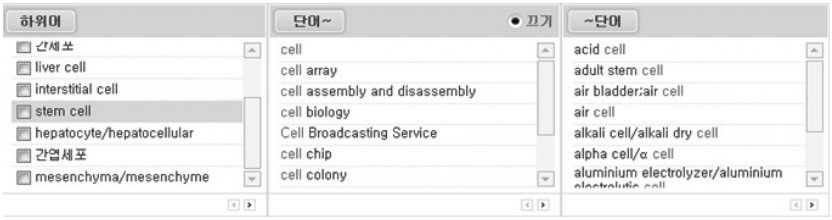

Fig. 8. Query suggest function with word intelligent network

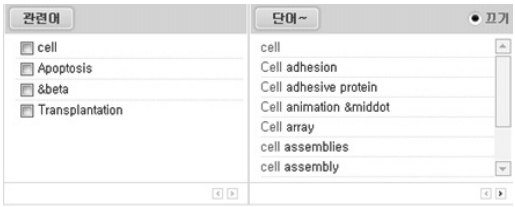

Fig. 9. Query suggest function with topic network

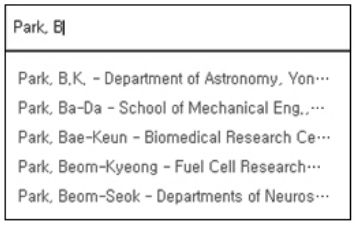

Fig. 10. Query suggest function with authority data

4.2. Query expansion functions

In the terminology dictionary and the word intelligent network, various semantic relations and Korean-English translation information is included, and using this information, semantic query expansion functions were implemented to enhance the search efficiency and the accuracy of the search results.

The inconsistency issues between the index and search terms were sought to be resolved by automatically expanding and searching queries, utilizing the synonym information of the word intelligent network, and the headings and variants information of the authority data, and it was therefore possible to search for related documents with higher accuracy. Also, translation information and abbreviation information of the terminology dictionaries were added to the query expansion searches to provide the word-based cross-language information search functions, to search for both Korean and English documents, and to decrease the repeated search frequencies of the users.

4.3. Utilizing the identifying information

The additional information for the identifiable entities proposed in this study are authority data identity IDs, topic networks, co-author networks utilizing word intelligent networks, annual research trend graphs, and the research topic networks. The authority data identity ID may be brought back from the authority DB, and through this ID, and referencing the authority data mapping table, a document ID corresponding to the identity ID can be obtained. Using the document ID list referencing the corresponding author IDs gained from referencing the academic journal authority data mapping tables, an author network as shown in Figure 11 may be created.

Also by referencing the annual information for the document ID lists stored in the academic journal DB, annual research trend graphs may be created, and by referencing the keyword information for the document ID list stored in the academic journal DB, research topic networks as shown in Figure 12 may be created.

The academic information services utilizing this identifiable information are useful for providing research trends and expert, social network services to users.

Fig. 11. Example of author's network

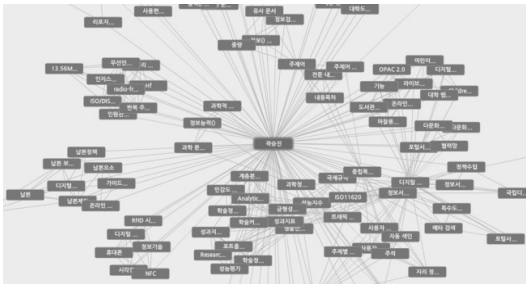

Fig. 12. Example of topic network

4.4. Meaning discrimination technology

The terms may be discriminated by meaning for each defined fields. For example, when the keyword ‘cell’ is searched in ordinary search systems, the results list a search of web documents including the word ‘cell’. But the word ‘cell’ stands for data transmission and storage units of the network in the computing fields, an equipment for generating energy in the electric and electronic fields, and a basic unit of living organisms in the medical fields, and has different meanings for different fields. The pre-existing search services provide documents including the input query without distinction of field, but the service technology proposed in this study can discriminate between the different meanings of the word ‘cell’ used for wireless communications in the computing fields, used as batteries in the electric and electronic fields, and used as cells of living organisms in the medical fields. The word intelligent network includes the many meanings of the word ‘cell’ and the relevant translation information, and by utilizing this discriminated information may be provided to the users, giving this technology the advantage of providing the expert information from the wanted fields for the users.

To apply the meaning discrimination technology, an analysis of the user trends for the service should be performed in advance. In other words, development of analytic techniques such as search pattern analysis of experts in the field of science and technology, expert trends analysis, and related topic analysis, should be done first, to provide core technologies for the expert-centric knowledge base system.

5. Evaluation of usefulness

5.1. Evaluation methods

In this study, an academic information service prototype, with the academic information service functions proposed in Section 4 applied, was implemented and compared to the DBpia, a Korean academic database. The evaluation items were “usefulness of the query suggest functions”, “usefulness of the query expansion functions”, “usefulness of the identifiable information utilizing functions”, and “usefulness of the meaning discrimination technology”, and a hypothesis stating that “there is

a difference in the usefulness between the knowledge base academic information service and the ordinary academic information service" was established.

For the analysis of the evaluation results, the collected data from a total of 30 people was used as samples by the SPSS version 12.0 statistic packages, and was statistically analyzed. To test the hypothesis, a paired-samples t-test was used to test the difference of average between groups for each corresponding pair, to perform a comparison analysis on the search efficiencies between the ordinary system and the experimental system, and the level of significance (p) used for these was 0.05.

For evaluating the usefulness of the ordinary academic journal service functions and the knowledge base academic journal service functions, a questionnaire was given to the test subjects and the average values of the 5-point Likert scale, 1) strongly disagree, 2) disagree, 3) neither agree nor disagree, 4) agree, 5) strongly agree, were used. To perform comparison analysis of the usefulness, the hypothesis was tested using the questionnaire items as variables.

5.2. Evaluation results

From the t-test, the results showed significant differences between the usefulness of the two systems, with the t-value and p-value all under the level of significance of 0.05. In particular, for all items, the knowledge base academic information search system was shown to be superior to the ordinary system. It showed unequaled satisfaction over the ordinary system for providing knowledge base query suggest functions, query expansion functions, grouping of identified users for search results, and visualized research activities information related to the author, to the user Therefore, it is considered that if this technology was used not only for information management, but also as a tool for information searching in the future, maximization of user satisfaction may be possible.

Table 1. Evaluating the usefulness of academic information services

Evaluation item	System category	N	Average	Standard deviation	Mean difference	t-value	p-value
Query suggest function	Comparative system	30	3.533	0.89	-0.7	-3.23	0.001
	Proposal system	30	4.233	0.77			
Query expansion function	Comparative system	30	3.466	1.008	-0.44	-1.676	0.049
	Proposal system	30	3.9	0.994			
Identifiable information utilization	Comparative system	30	3.933	0.827	-0.33	-1.69	0.047
	Proposal system	30	4.266	0.691			
Meaning discrimination technology	Comparative system	30	3.5	1.008	-0.73	-4.025	<0.0001
	Proposal system	30	4.233	0.994			

6. Conclusion

In this article the authors described the knowledge base necessary for intellectualization of science and technology information, and evaluated the usefulness of the intelligent information management and service model of science and technology academic information utilizing this technology. Through the knowledge information management and service system based on knowledge bases, it was confirmed that not only the user but also the manager could use this system to configure ways to manage and service precise and meaningful information.

Through the system utilizing the knowledge base, users can find knowledge they did not expect to find, and create and share new knowledge. Of course, for the establishment and meaningful interfacing of the knowledge base, a great deal of manpower and time must be invested, and so it is also very important to plan ways to allow maximum automation of the knowledge base establishment process.

Recently, various methods to allow systemic management, analysis, and provision of meaningful information to the users and manager by application of semantic technology have been proposed, and among these, studies on the linked data concept based on semantic technology have been actively carried out. Linked data is being recognized as a new method of disclosing, interfacing, and sharing data, by providing a distinguishable name that can access the semantic data web, especially an URI that can dereferenced, and describing the semantic data in a triple structure through RDF, publishing the data using HTTP URI and establishing relations with each other, to allow interpretation/re-interpretation and use/reuse by human and machine. Therefore, as the structures of the knowledge base in this article are composed based on identifying codes similar to connection data, it is predicted that the knowledge base can be expanded to connection data, giving it huge functional capabilities. In the future, studies including these aspects will be performed.

References

Arampatzis, A. T., T. Tsoris, & C.H.A. Koster. (1997). IRENA: Information Retrieval Engine based on Natural Language Analysis, *In Proceedings of RIAO97 Computer Assisted Information Searching on Internet*, 159-175.

Brachman, Ronald J., & Levesque, Hector J. (2004). *Knowledge Representation and Reasoning*. Elsevier, Inc.

Choe, Ho-Seop. (2006). Construction Method of Large-scale 'Urimal(Korean)-Word Intelligent Network'. *Hangul*, 273, 125-151.

Choe, Ho-Seop, & Ok, Chul-Young. (2004). Information Retrieval System and Ontology. *Communications of the Korean Institute of Information Scientists and Engineers*, 22(4), 62-71.

Evans, D., & Zhai, C. (1996). Noun-Phrase Analysis in Unrestircted Text for Information Retrieval. *Proceedings of the 34th Annual meeting of Association for Computer Linguistics*, 17-24.

Im, Ji-Hui, Choi, Ho-Seop, Bae, Yeong-Jun, Ok, Cheol-Yeong, Choi, Seong-Pil, Seong, Won-Gyeong, & Park, Dong-In. (2005). Construction of immunology thesaurus and ontology. *Annual Conference*

on Human and Language Technology, 2005, 21-27.

Park, Jung-Oh, & Hwang, Do-Sam. (2000). A terminology extraction system. *Proceedings of The 27th KISS Spring Conference 2002*, 27(1-B), 381-383.

Rowley, J., & R. Hartley. (2008). *Organizing Knowledge: and Introduction to managing access to information*. Burlington.VT: Ashgate.

Seung, Hyon-Woo, & Park, Mi-Young. (2003). A clustering technique using association rules for the library and information science terminology. *Journal of the Korean Society for Library and Information Science*, 37(2), 89-105.

What is CYC? Cyccorp, INC. <http://cyc.com> [cited 2013.03.05].

Zins, Chaim. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58(4), 479-493.