情 報 学 基 礎 35－4
（1994．11．9）

概念中心型文書管理と全文検索による情報共有

藤澤 浩道，加藤 寛次*，小島 啓二*，友広 修造**，和歌山 哲**

(株) 日立製作所中央研究所
*(株) 日立製作所システム開発研究所
**(株) 日立製作所ソフトウェア開発本部

グループワークにおける情報の共有化には，事実やデータといった具体情報の共有と，仕事の世界観の共有という2面があることを述べる．この内後者については，これまであまり着目されないが，概念を中心に捉える文書管理の方法を提案する．前者については，全文検索が有効なことを述べる．日本語文書の全文検索には，シソーラスを用いて自動索引付けを行う方法もあるが，索引付けが不要なシーケンシャルサーチを基本とする方式について述べる．これによってシソーラスからもれることが多い新語や固有名詞についても検索漏れのない信頼性の高い検索が実現できる．

Concept-centered Document Management and Full-text Search for Information Sharing

Hiromichi Fujisawa, Kanji Kato†, Keiji Kojima†, Shuzo Tomohiro†† and Satoshi Wakayama††

Central Research Laboratory, Hitachi, Ltd.
†Systems Development Laboratory, Hitachi, Ltd.
††Software Development Center, Hitachi, Ltd.

ABSTRACT

A new approach to the document information management is presented. It is aimed to achieve information sharing in an organization from two aspects that are conceptual work space sharing and facts-and-data sharing. Instead of managing documents in hierarchical folders, it manages a conceptual space represented as classification hierarchies of "concepts" where documents are positioned. Users interact with the displayed representation of the space to access and store documents, resulting in common understanding of the conceptual work space. A fast sequential full-text search method for Japanese documents allows high recall and high precision document retrieval which is important to achieve "on-demand" type facts-and-data sharing.

Keywords: information sharing, document management, information retrieval, conceptual work space, multiview classification, full-text search, synonym generation, spelling variants

1. Introduction

Document information systems are now entering a new age of network-based, corporate-wide document processing. Most of PCs are now being networked, and used for word processing and communications. It is said that more than 80 percent of electronic corporate information is in the form of documents, and remaining part is structured record data. The document is literally a knowledge and information representation medium, and hence an indispensable medium for human to human communication.

The document information system is therefore one of the most important basics in the office information systems. A remarkable increase in the knowledge workers' productivity is expected. The key to this is "information sharing" in the workgroup members by the use of such systems. However, it is not sufficient just to have centralized or distributed document management servers by which the group members could see the same set of documents. It is difficult for them to identify important messages buried under too many documents, to retrieve right documents even from a vague memory, and so on.

We discuss requirements to achieve better or real information sharing, and present a document information management system that features methods of concept-centered document management and fast full-text search. Index-free full-text search of Japanese documents has waited for a technical breakthrough, and we have developed a two-stage surrogation method for fast multiple string search to realize real full-text search of Japanese documents. The method is now applied to our software product, Document Information Management System, *Bibliotheca* (originally a Greek word meaning "library"), which is a client/server type system running on PCs and UNIX workstations.

2. Requirements for Information Sharing

A direct interpretation of "information sharing" would be the status of knowledge workers to see the same information or to be able to do so. It is believed that it is the document management system that can realize it. This kind of the system provides document database servers on a local area network and client software running on personal computers, by which users at different locations can see the same documents in these document database servers. It is therefore possible for them to share information in theory. But it is not so simple in reality because of human factors and other factors.

There are two aspects in information sharing. The first one, which is very straightforward, is with regard to sharing of facts and data related to the work, which are conveyed by documents most of the time. This is the aspect of the information sharing in a normal sense. However, it is hardly possible for an ordinary person to see every document every time when it is created or updated by his/her group members, especially when the group is big. There are too many documents created and updated by the members. Important information might be buried under less important documents. Of course, people may send an e-mail message telling other members which documents are important and how important they are. But, some of the documents may become important later in other contexts. Therefore, they have to rely on the "on-demand" type accesses to the documents. Here are the human factors and technical issues around the information retrieval. The access patterns of the on-demand type are ad hoc and non systematic, requiring free term search in the full text.

As regard to information retrieval, it is also important to note that documents never exist alone, and that document traversal by following associative links (hypertext links) is an indispensable function.

The other aspect of information sharing is the common understanding of the "conceptual work space" they are involved, which we believe is as important as the facts-and-data sharing. Smooth communication, being one of the objectives of information sharing, requires a common vocabulary and a common view of that space. The conceptual work space is defined to be a represented form of showing what kinds of the things exist associated with the work, what they are, and how they are related to each other. It is a kind of knowledge required to carry out the job actually. It is the knowledge about organizational structures, projects, technology, developments, products, customers, public relations, many kinds of forms and documents, and so on. Unless people understand these things, no communication is possible as a matter of fact. This issue has not

been targeted well by conventional document related systems.

There are other issues as well which are orthogonal to these aspects. Groupwork is an aggregation of individual work. Therefore, both the individuality and conformity need to be taken into account. For instance, individuals may see the work space slightly differently from each other because of his/her different specialty and deepness of thought. Therefore, personal views to the work space should be also allowed in the system. Furthermore, information and documents an individual is processing during the work are fragmental, piece-wise, and rather disordered usually. As a result of their work, they create and output well-organized information in terms of documents which are going to be public in the group. Supporting this process of organizing a collection of piece-wise information into well-structured information is also important.

This paper discusses the two aspects of the conceptual space sharing and the facts-and-data sharing. For the conceptual space sharing, the approach presented in this paper is a document management system that enables users to see their documents from the window of the conceptual work space. We call this approach "Concept-centered Document Management." The conceptual space is represented in terms of "concepts" and "relations" to which documents are associated, rather than concepts are associated to documents. Users may select relevant concept nodes and see the associated documents. Because the conceptual space is a window to the document space, users' memory of how things are organized is refreshed each time when he/she stores documents and/or makes an access to them.

For the facts-and-data sharing, index-free full-text search is presented which is to enable them to make flexible on-demand type accesses to the shared documents. Full-text search itself has not been so special in the case of English documents, but that of Japanese documents has been waiting for a technical breakthrough. The presented method of full-text search adopts sequential scanning of flat text files, and a two-stage surrogation method to speed-up the search process. It features automatic generation of equivalent expressions of query terms including synonyms and spelling variants. At most, 1024 terms can be searched in a single scan without additional search time, realizing very fast, highly reliable search. This method has solved the inherent problem in the automatic full-text indexing that relies on a thesaurus. It is the problem that words that are not listed in the thesaurus cannot be indexed. It is not a surprise that a name of a person who has become famous recently is not listed, for example.

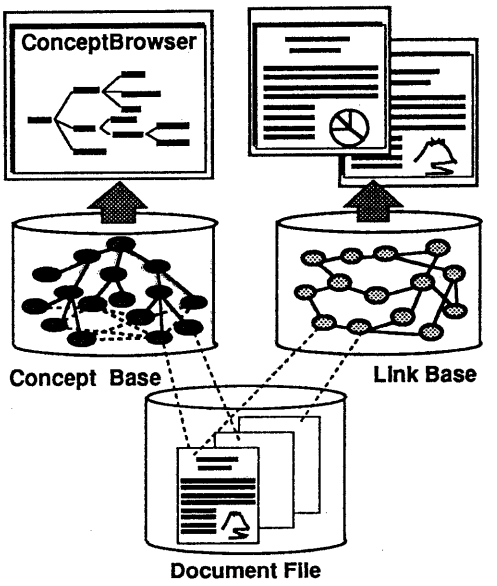

Fig. 1 Conceptual diagram of concept-centered document management

3. Concept-centered Document Management

Concept-centered Document Management is a new approach to the way of showing users the document space. Conventionally, documents have been at the center of management and hence, users think of documents first. Attributes such as file names, authors, keywords, classification codes, etc. are just an attachment of a document. Consequently, the attached attribute data are independent each other and are not the objects to be managed. The method could be called "document-centered" in that sense.

The proposed approach is, on the contrary, to provide users with a representation of the conceptual space to which documents are associated. The image of the approach is shown in Fig. 1. ConceptBrowser is a software program which manages a representation of a

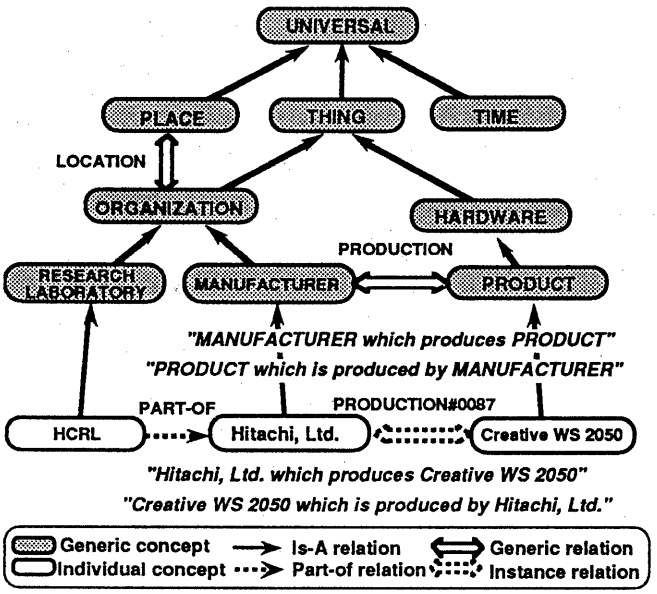

Fig. 2 Concept-relation model

conceptual space in terms of classification hierarchies of thing's names or concepts. It allows users to browse through a network of concepts, and displays documents associated with a selected concept as requested. It may also have a link base to store hypertext links to enable hypertext browsing.

We developed a research prototype ConceptBrowser based on a concept-relation model [1], [2], and conducted test use experiments. The concept-relation model is a kind of semantic networks [3], [4], but it discriminates generic concepts from individual ones, and likewise, discriminates generic relations from instance ones as shown in Fig. 2. The subnetwork of generic concepts and generic relations defines a schema or a meaning structure of concepts. By doing so, the generic framework of knowledge can provide intelligent prompts for user's editing and querying interactions. Concepts may have multiple parents when it has multiple aspects. For instance, "laboratory" is an "institution" for research, a "location" where researchers work, and a "building" we can see. Important facts and thing's attributes can be represented in terms of individual concepts and instance relations. They are displayed on the screen in a sentence or a table.

By using its network editing functions, we registered about 450 computer-related news articles with associated concepts to see what kind of and how many generic concepts and relations were necessary. As a result of such registration, the concept network come to have, approximately, 3,500 concepts and 7,000 relations, which included 600 generic concepts and 90 generic relations. Generic concepts included those representing people, organization, location, time, event, domain specific concepts, and documents.

The experiments showed the effectiveness of this approach especially for the "evolutional information processing." Namely, it was quite easy to refine, update, and restructure the classification. The office work of information collection, understanding, and generation of new information in terms of documents is an evolutional process.

Based on the study, a software product, Document Information Management System, Bibliotheca/IS (InfoShare), has been developed. The software system consists of a document information management server running on Unix workstations and a client program running on a PC with Microsoft Windows 3.1. The server software is implemented on the object-oriented database engine, Bibliotheca/OM (Object Manager).

It manages conceptual spaces in terms of concept hierarchies, and documents with their attributes which include bibliographic items, security control attributes, related concepts, relational links to other documents, and other user-defined attributes.

One of the most important features of this system is "Multiview Classification" which is based on the orthogonal classification hierarchies as shown in Fig. 3. This feature is the essence of the approach studied in the ConceptBrowser research. The client software enables users to make "multiview browsing" and select any number of concept nodes to confine a subset of documents. The GUI appearance of Bibliotheca/IS is shown in Fig. 4. What is important here is that documents associated to the subsumed concepts of a selected concept are also retrieved due to its inheritance mechanism. Users can retrieve

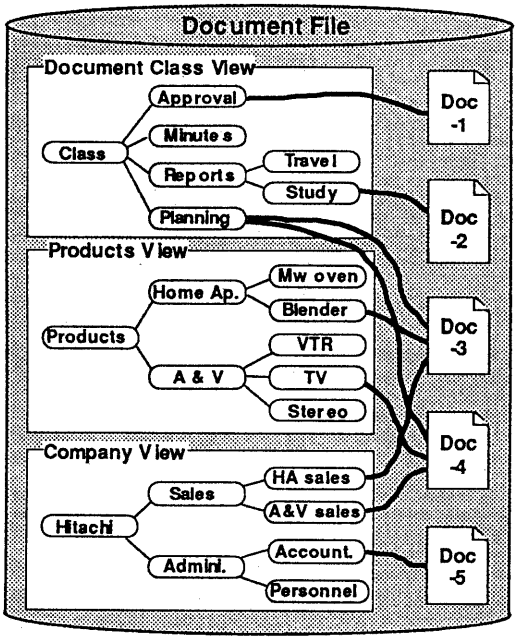

Fig. 3 The concept of Multiview Classification

desired documents even from a vague memory, because they can specify more general concepts that they can remember. For example, instead of specifying very concrete concept, say "microwave oven", he/she may select a superconcept "home appliances."

Classification hierarchies can be created and edited by end users as in ConceptBrowser so that the classification can be adapted to follow evolutional changes during the work. The classification data resides on the server, and users automatically share the same multi-classification views. The rights to make changes can be controlled by the access privileges. It is a future problem to manage both personal views and a common view in an organized manner. One technically difficult problem is the way to distribute and to merge different views.

It may be argued that the conventional desktop metaphor showing a hierarchical file system as classified folders would also act as a representation of the conceptual work space. However, it is not true. The hierarchical file system has intrinsic problems. Namely, because it is based on a single hierarchy, the classification system becomes confused, and branches near leaves of the hierarchy are duplicated when things have multiple viewpoints, which is always true. Multiview

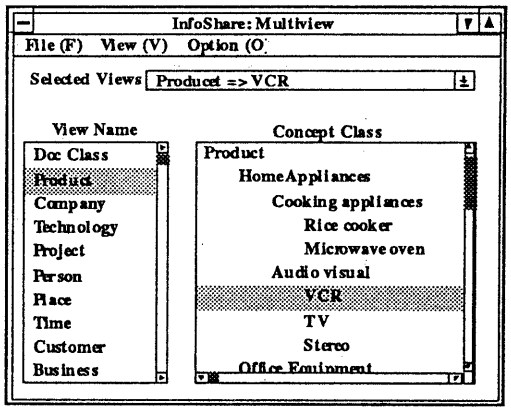

Fig. 4 GUI of a Bibliotheca/IS client

Classification may have as many hierarchies as needed, each of which can be "orthogonal" to each other.

4. Full-text Search of Japanese Documents

In addition to the classification approach, an easy-storage-easy-retrieval approach is also important. Although the classification is important as a view to a work space, it is also true that spontaneous, context-free information retrieval can take the real value out of the share of the documents. By borrowing others' knowledge and experiences, work can be accelerated.

One of the solutions is the full-text search. It is content-based information retrieval without requiring human indexing. Users can pose any query terms to the system. In this sense, it is context-free. Query terms can form a complex condition such as Boolean logic, proximity condition, contextual condition, and their combinations.

Although the full-text search of English text can be realized rather straightforwardly by indexing every word appearing in documents, the same method cannot be applied to Japanese text. Japanese text does not have delimiters such as spaces as in English. To delimit words, lexical analysis of the text using a thesaurus is required. Hence, only the words listed in a thesaurus can become indexes. The problem of this "automatic indexing" means that person's names, company names, newly coined words, abbreviations, and so forth are hard to be indexed, killing the real merits of the full-text search. It often happens that one day, an unfamiliar name becomes very important.

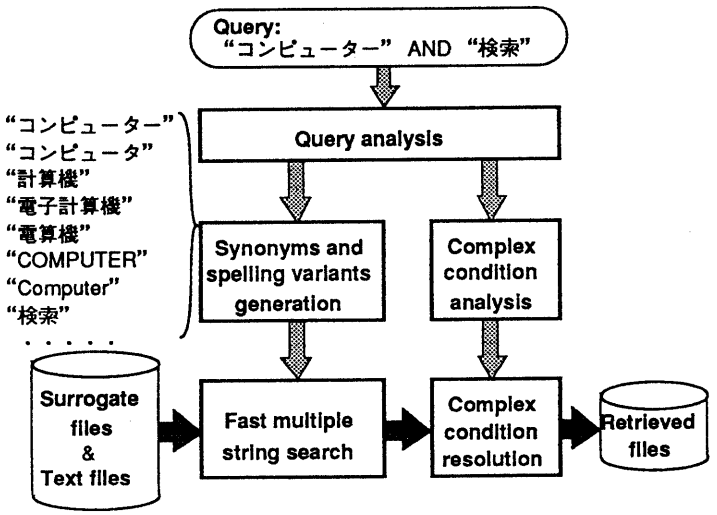

Fig. 5 General flow of the full-text search

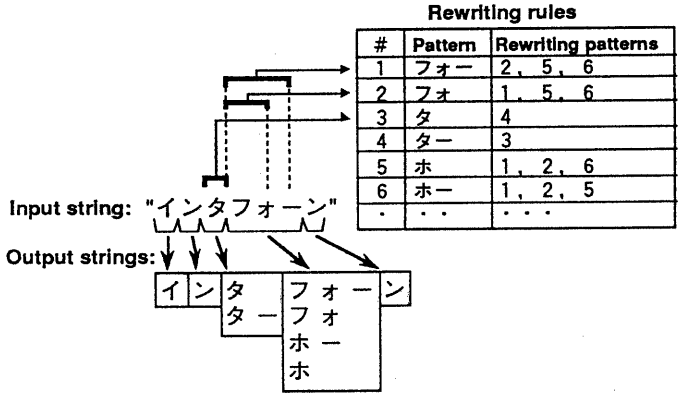

Fig. 6 Spelling variants generation

A new method for full-text search of Japanese documents presented here solves these problems and also the problem of synonyms and spelling variants which has not been treated well so far. As shown in Fig. 5, a complex query expression is first analyzed, and appearing terms are expanded to include synonyms and spelling variants. Synonyms are generated by a synonym dictionary, and spelling variants of Katakana words are generated by applying rewriting rules. More than 1,000 rules have been created to cover variations in Katakana expressions of foreign-borrowed words. Possible combinations of local alternatives in a Katakana string are generated as shown in Fig. 6. To search the generated multiple strings in a single scan, the sequential search method is based on a finite-state automaton which is an extended version of the Aho-Corasick algorithm [5]. At most, 1024 terms can be searched simultaneously.

To overcome the speed problem, we have incorporated the two-stage surrogation [6] shown in Fig. 7. The first stage surrogation is based on a bit map which consists of bit vectors. Each bit vector corresponds to a document, and shows which characters out of an eight-thousand character set appear in the corresponding document. Bit ONE in a vector means that a character that gives the bit address value by applying a hash function to the character code appears in the document. Therefore, search is done by finding bit ONE vertically at the bit addresses corresponding to the searching characters, identifying documents to be searched at the second stage. Because the bits are overloaded, it may include "noise" documents.

The second stage surrogation is based on a condensed text file, which is a residue of the exclusion of duplication and Japanese counterparts of "stop words," i.e. auxiliary words in the original text. String search through the condensed text file can answer the complex queries of Boolean logic. More complex conditions such as proximity and contextual conditions require the final stage of searching the original source text. The proximity condition is "two query terms to appear with (or within) a specified distance." The order of the two terms can be specified if necessary. The contextual condition is "two terms to appear in the same sentence or paragraph at the same time." The order can be specified in here as well. Further complex conditions are Boolean combinations of such contextual and/or proximity conditions.

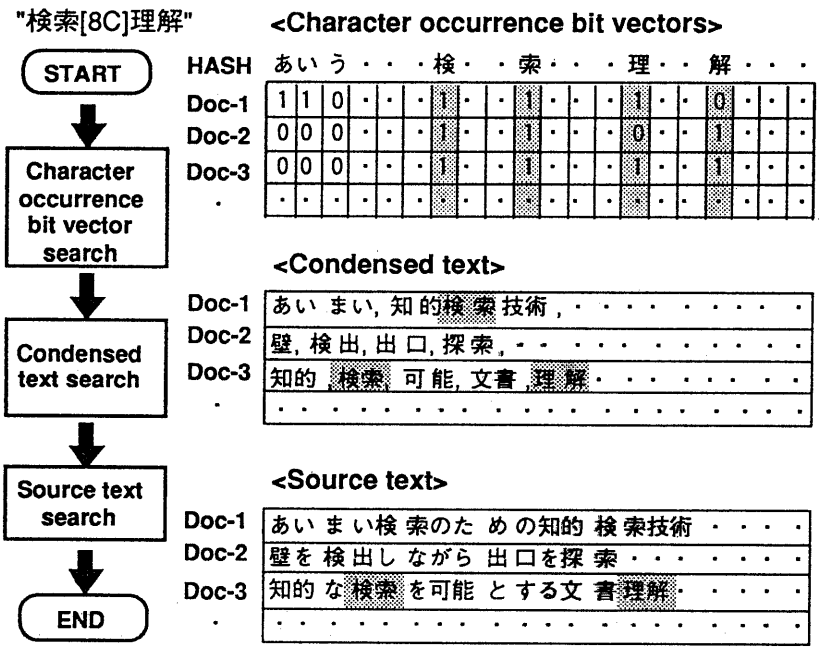

Fig. 7 Two-stage surrogation method

Table 1 General specification of Bibliotheca/TS

Items		Description
Search System	Documents	100,000 documents per database
	Databases	16 databases
	Functions	Full-Text Search/ Field Text Search Synonyms & Spelling Variants Generation Multiple Strings Search Boolean, proximity, context, range conditions
	Retrieval Time	Simple Query : 2 - 8 seconds (300～1,600MB)
		Complex Query : 2 - 20 seconds (300～1,600MB)
Search Engine	Algorithm	Modified Aho-Corasick Method
	Query Terms	Max 1,000 terms
	Scan Rate	Max 8.5 MB/s
Disk Storage	Capacity	3 - 12 GB
	Transfer Rate	Sequential: 8 MB/s Random: 2.5 MB/s

A hardware text search machine, TSM-I, was developed as an experimental prototype [6]. It had twelve small disks from which the source text data could be read in parallel, and a special hardware search engine which could scan a text string at the speed of 20 mega-byte per second for a multiple string search. By applying the two-stage surrogation method, the system search speed could reach 100 mega-byte per second.

Bibliotheca/TS (TextSearch) is a software version of the text search machine, which is a commercial product now. Users may retrieve documents stored in the server on a Unix workstation from a client running on a PC with Microsoft Windows 3.1. The retrieval time is

about 3 seconds for 100 thousand 1KB-long documents by using a 105 MIPS workstation, Model 3050RX/330. It varies depending on the complexity of the query and the hit ratio at each stage. The general specification is shown in Table 1.

5. Conclusion

Two aspects of information sharing has been discussed. It has been argued that rather than classifying documents, it is more important to classify things because that classification is a representation of his/her world view. The proposed Multiview Classification gives users the power to browse and navigate in the conceptual space and the document world as a result.

Index-free full-text search is another approach to information sharing. A new method for Japanese documents has been presented. It features automatic synonyms and spelling variants generation and very fast multiple-string search based on the two-stage surrogation. It supports user's spontaneous, ad hoc information accesses. We think that classification and search are two sides of the same coin.

The proposed method and system can elevate the users to the third step of the following four steps of information sharing;

- to have distributed or central common store of shared documents,
- to be able to search and retrieve documents from the common store easily,
- to be able to see and use the common classifications and common world view, and
- to be able to find persons easily who might need the information concerned.

The fourth step will be reached by automatic document classification and "information filtering" which is based on information retrieval techniques [7] and our system will feature this in a future version. Semistructured documents based on SGML (Standard Generalized Markup Language) will be also a key in the sharing of information in that, by using SGML, the system may extract document attributes automatically, create appropriate information flow, and make a right storage.

As for the future version, the concept hierarchies will be used as a front end of the full-text search. In stead of a conventional thesaurus, they can be used to expand query terms into equivalent expressions and more specific words including pronouns. Because the classification can be easily updated by end users, it can be a powerful tool to enhance the full-text search capability.

Acknowledgement

The authors wish to express their gratitude to H. Mori, Manager, Hitachi Seibu Software, Ltd. (originally at Computer Division of Hitachi, Ltd.), and T. Ishizuka, Manager, Software Development Center, for their invaluable leadership of the group concerned.

Reference

[1] H. Fujisawa, A. Hatakeyama, and J. Higashino, "A Personal Universal Filing System Based on the Concept-Relation Model," Proc. 1st Int. Conf. Expert Database Systems, Charleston, SC, 1986, pp.31-44.

[2] H. Fujisawa, I. Kiuchi, T. Koguchi, and H. Kondo, "A Visual User Interface for a Personal Information Base using a Concept Network," Proc. Int. Symp. Database Systems for Advanced Applications, Tokyo, April 1991, pp.69-78.

[3] J. F. Sowa, "Conceptual Structures: Information Processing in Mind and Machine," Addison-Wesley, 1984.

[4] R. J. Brachman and J. G. Schmolze, "An Overview of the KL-ONE Knowledge Representation System," Cognitive Science, Vol. 9, 1985, pp.171-216.

[5] A. V. Aho and M. J. Corasick, "Efficient String Matching: An Aid to Bibliographic Search," Comm. ACM, Vol. 18, No. 6, 1975, pp.333-340

[6] K. Kato, H. Fujisawa, H. Kawaguchi, A. Hatakeyama, et al., "An Index-Free Full-Text Search for Large Japanese Text Bases," Proc. Advanced Database System Symposium '89, SIGDBS, Information Processing Society of Japan, Kyoto, Dec. 1989, pp.75-82.

[7] N. J. Belkin and W. B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?," Comm. ACM, Vol. 35, No. 12, 1992, pp.29-38