Text Structure Analysis as a Tool to Make Retrieved Documents Usable

Noriko Kando

Research and Development Department
National Center for Scientific Information Systems (NACSIS), Japan
kando@rd.nacsis.ac.jp

Abstract

This paper describes an information retrieval system with the function to support user's use of the retrieved documents using the text-level structure of documents. The text-level structure of each document is described by the occurrence of typical functional components in the text. Automatic detection of the components has been attempted in previous works using surface-level language processing. The proposed system firstly utilizes the text structure to conduct high-precision searches of documents or passages by distinguishing the role or function each concept plays in the text. It also allows browsing or skimming of retrieved texts, creating summaries on-the-fly with various levels of condensation specified by the user. Moreover, the system can search and display any unit of a text such as a sentence, a paragraph or a chapter. Comparison of relevant passages in retrieved documents across multiple texts is helpful for users to examine, analyze, compare and integrate texts and undertake such information work as decision making, problem solving or writing papers, etc., based on the information obtained from retrieved documents.

1 Introduction

One of the lines of investigation that we should pursue in information retrieval is the provision of enhanced functionality of post-retrieval processing to support the use of retrieved documents. As Strzalkowski mentioned in his recent book, a summary report, a list of facts, a one-page brief, or a chart are examples of such techniques for organizing and presenting the information in retrieved documents in a form that is immediately usable, rather than merely providing ranked lists and pointing to relevant "texts" as traditional IR systems do. (Strzalkowski, T., 1999) In particular, such "post-retrieval processing" to make retrieved documents more usable is needed in interactive information retrieval systems such as Internet search engines.

This paper describes an information retrieval system providing such post-retrieval processing functions. The characteristic functions of the system are as follows:

(1) To support the user's skimming or browsing in the text by highlighting the salient parts of the text according to the text structure or by structuring the text with category label tags that represent the text structure or the role or function the sentence plays in the text. The level of "saliency" can be specified by the user.
(2) To create the text's gist on-the-fly with various levels of condensation specified by the user.
(3) To utilize the text structure to conduct high-precision searches.
(4) To search and display any units of texts, from sentences, paragraphs or chapters to full-length texts.
(5) To support comparisons among the relevant sentences in the retrieved documents across multiple texts.

The following are the characteristic retrieval functions of the proposed system:

(1) Two versions of user interfaces: English and Japanese.
(2) Either interface can process both English texts and Japanese texts.
(3) Language-oriented indexing and query processing techniques (Kando *et al*., 1998a).
(4) Enhanced phrase processing based on linguistic knowledge (Kando *et al.*, 1998a).

Compared to the previous version of the system reported in (Kando 1997a), these points are improved:

(1) Acceptance of natural language sentences as queries (Kando *et al.*, 1998a).
(2) Changes to the initial interface to accept a user's query based on the survey of users' behavior reported in (Kando 1997c).
(3) Changes to the algorithm for automatic detection of text structure and to the retrieval algorithm, so that the system can perform effectively against texts in which only some sentences contain category labels. The previous version presupposed that every sentence contained one or more category labels.
(4) Improvements in ranking algorithms (Kando *et al.*, 1998a).

Manual query construction (Kando, 1997a) meant that the problem of automatic segmentation of Japanese natural language queries into search terms could be avoided. However, constructing queries using operators and category labels, as shown in Figure 1, is much too complex for users. As a comparison of three user interface models for query

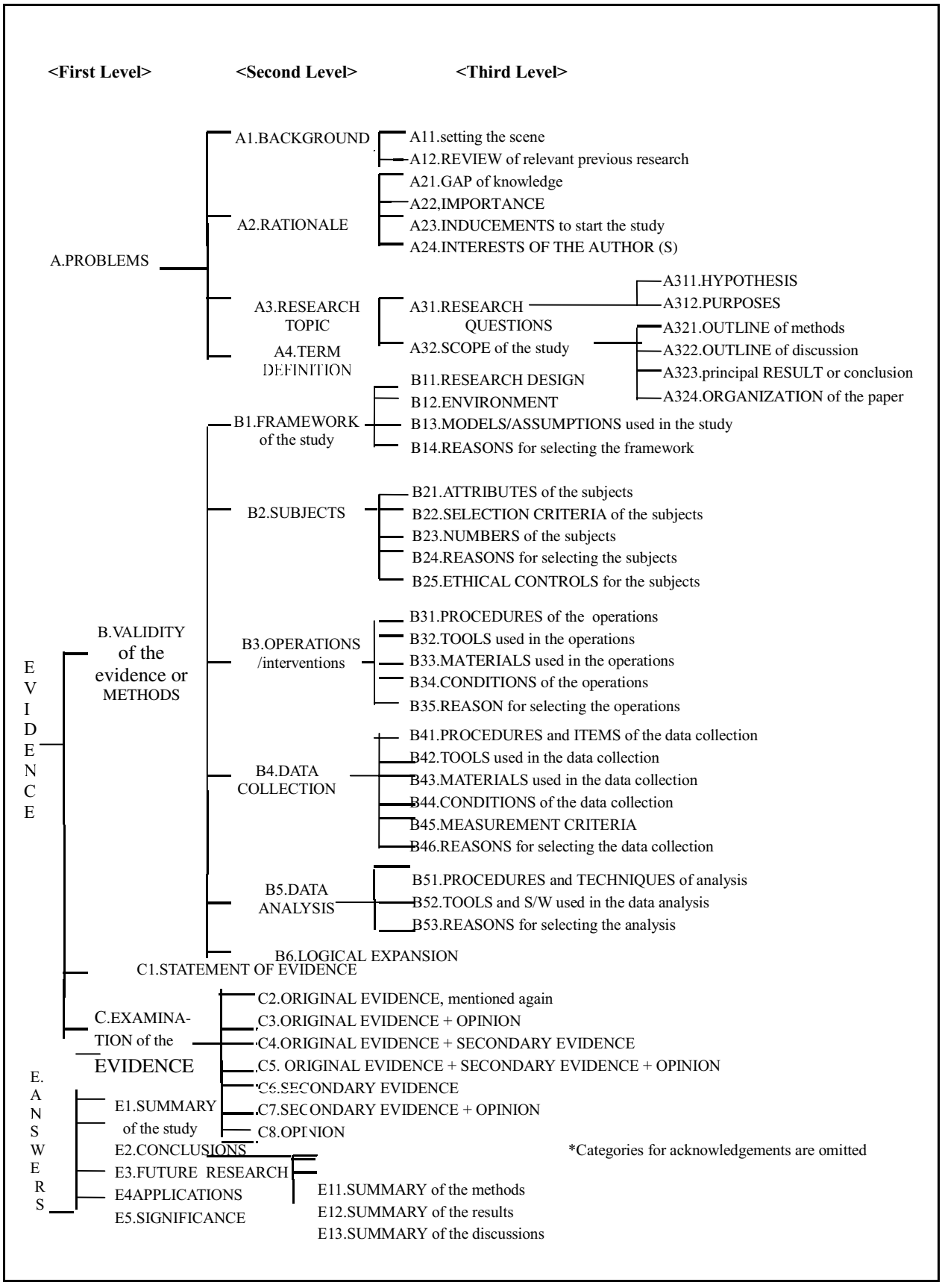


Figure 1. The categories

construction using text-level structure (Kando, 1997c) suggests, improvement of the interface is indispensable.

Although IR systems have focused on the phenomena of term occurrence or sentence-level structure, the characteristic aspects of the text and the text type should be incorporated in text based information-processing systems. Among the various approaches to text, this paper focuses on the typical text structure of a genre, or the text type[1]. It is known that each genre has a typical structure of text. Many researchers have shown that research papers and their abstracts possess a highly typical structure (Jones, 1995; Kando, 1991; Liddy, 1991; Rama and Srinivasan, 1993; Samuels *et al.,* 1988; Swales, 1990). Such structure can be described with a set of typical components and their order in a text (van Dijk, 1980). "Background", "reference to the previous research", "purpose", "methods", "results", "discussion", and "conclusion" are examples of typical components of the genre of research papers. A more detailed structural scheme is used in this paper (Figure 1). It is a natural structure of the informational content of the text that is familiar as a kind of social convention to the people who use text of that genre to communicate. It is therefore expected to be usable by users in a search process (Oddy *et al.*, 1992).

The text-level structure of research papers has been applied to various areas such as: indexing (Fuller, 1984), automatic abstracting (Oakes and Paice, 1999; Lehman, 1999; Teufel and Moens, 1997; Endres-Neggemeyer *et al.*, 1995; Jones, 1995; Paice and Jones, 1993), reference interviews (Allen, 1988; 1989), text retrieval (Kando, 1997a; Liddy, 1991; Oddy *et al.*, 1992; Rama and Srinivasan, 1993), browsing in a text (Miike *et al.*, 1994), information extraction (Oakes and Paice, 1999; Jones 1995; Paice and Jones, 1993), and the design of user interfaces and electronic journals or digital library systems (Bishop, 1998 and 1999; Dillon, 1994; Miike *et al.*, 1994). While this paper focuses on the genre of research papers, the principles of text-level structure apply widely. In particular, information extraction and automatic summarization using text-level structure has been attempted for various types of text (Jang and Myaeng, 1997; Moens and Uyttendaele, 1997).

The next section discusses the problems related to information retrieval and clarifies the approaches used in this paper. Section 3 presents the structural scheme used here; Section 4 describes the methods and results of experiments. Finally Section 5 discusses the implications of the approaches proposed and indicates further possible studies.

2 Structural Schema of Research Papers

My colleagues and I have conducted a series of analyses on text-level structures. We have delineated the typical structure of research papers (Kando, 1991), English and Japanese newspaper articles (Kando, 1995a; 1995b; 1996), Nursing diaries (Kando, 1995a) and photographs and videotapes of TV news (Ueda *et al.*, 1995). We have studied automatic detection of structure (Kando, 1997b), shown that searches using text-level structure are more effective than ones that do not distinguish the role or function each component plays in the text, and suggested various applications (Kando, 1997a).

The full set of categories used in the system is presented in Figure 1. Each category represents the role or function that the component plays in the text, the relationship between a component and the whole text, and the relationship between components in the text and other components in other texts.

The categories are arranged hierarchically. The ***leaf categories***, which are the most specific level of categories, are assigned to each sentence. More than one *leaf category* label can be assigned to a sentence. The *leaf categories* can be translated into an upper level category via hierarchy links when the overall structure of the text is described. The characteristics and definitions of upper level categories are inherited by specific leaf categories through the hierarchy links. This inheritance is also used in the rules for automatic detection of the categories.

Text-level structure in each research paper was described by the category occurrences and their order in the paper. The structure was more complex than the "IMRD", or "Introduction, Methods, Results, and Discussion" organization, which is frequently mentioned as the structure of research papers in writing manuals (Day, 1991). The patterns of category occurrence in a text were categorized. The detailed list of the patterns was reported in (Kando, 1991).

Regarding automatic category detection, three kinds of rules, *i.e.,* ***(a) indicative clue phrases, (b) category order,*** and ***(c) scope of components*** were constructed manually and used in these studies. The rules (a) are essentially templates. They may include groups of indicative clues or phrases, the category of the component preceding the one being categorized in the paper, a position in an article and a paragraph, and the presence or absence of citations. Indicative clues were selected manually from training corpora. Similar words and phrases were also selected from writing manuals and thesauri, then formed into "***groups of indicative clues***". They are basically semantic categorizations of indicative clues or words in the clues. The term "***component***" refers to a series of one or more sentences of the same category. The rules (c) identify the successive sentences in a component, mainly based on connectives and anaphoric expression. Some of the rules provided stronger evidence than others, so each rule was assigned a weight. The rules (b) were based on the probabilities with which categories succeeded each other in the training corpus.

Based on this, the next section describes the results of the experiments and the characteristic functions the system implemented using a structure-tagged full-text database.

3 The Proposed System

3.1 Configuration

The system was implemented as a client-server environment. The server machine environment consisted of a WWW server (any kind can be used, our experimental system currently uses apache-httpd), the cgi developed for the system, a Japanese morphological analyzer, an English parser, and OpenText Ver.6 as a search engine. The client can use any operating system that can process two-byte character codes, and a Web browser.

3.2 Database

The experimental full-text database consisted of a few hundred Japanese research papers on Viral Hepatitis type C. The full texts of papers were scanned and OCRed after obtaining permission for their use in the experiment from the publishers.

Tags for the categories, which are shown in Figure 1,

[1] It is sometimes described as the "text grammar" approach, for example (Moens and Uyttendaele, 1997; Rama and Srinivasan, 1993).

and the components of logical structure of documents, *i.e.,* <article>, <title>, <sec> for section, for paragraph, etc., and the scores showing the belief value of the category assignment were assigned automatically. The components of author, authors' affiliations, captions for figures and tables, lists of references are excluded. An example database record is shown in Figure 2.

<article><title>非A非B型慢性肝炎に対するインターフェロン療法</title>
<body><sec><h1>はじめに</h1>
<s><A11><score>0.89</score>非A型非B型慢性肝炎に対するインターフェロン(IFN)療法が試みられるようになり、血清トランスアミナーゼの低下や組織学的な改善が報告されている。</A11></s>
<s><A21><score>0.54</score>しかし4週間の短期連日投与では，投与中止後血清トランスアミナーゼは再上昇を示すことが多く，投与方法になお若干の 問題点が残されている。 </A21></s>
<s><A312><score>0.93</score>今回筆者らはより有効なIFNの投与方法を検討する目的で，①連日投与と間歇投与の有効性の比較検討，

[English Translation of an Example of above]

<article><title>Interferon Therapy of Non A Non B Hepatitis</title>
<body><sec><h1>Introduction</h1>
<s><A11><score>0.89</score>It has been reported that interferon (IFN) treatment declines the serum transaminase level and improves the histological condition.</A11></s>
<s><A21><score>0.54</score>However there are some problems in the methods of administrationcsince many cases whose level of serum transaminase increased after continuous administration of IFN for four weeks were reported.</A21></s>
<s><A312><score>0.93</score>In order to discuss the more effective method of IFN administration, The authors conducted

Notes: A11: Setting the scene, A12: Review, A21: Gap of knowledge, <Article>: article, <title>: title, <body>: body of text, <sec>: section, <h1>: section heading, : paragraph, <s>: sentence

Figure 2. An Example of the Experimental Full-text Database

3.3 The Search Engine

The search engine is OpenText6 (OpenText Corp., Canada), which has a strong search function for structure-tagged databases. Any part of text enclosed by a beginning tag <tag> and an ending tag </tag> is called a *region*. It is a unit for a search and returned sets. The search engine can process hierarchically nested or overlapping tags. "Including" and "within" operations among regions are available. With this engine, complex queries with structural relationships can be processed, for example, "any articles in which the term 'rat' occurred in the components of the categories under 'B2 (Attribute of subject)' ", or "any paragraphs in which the word 'age' occurred in 'category C1 (statement of original evidence of the article)' ", and so on.

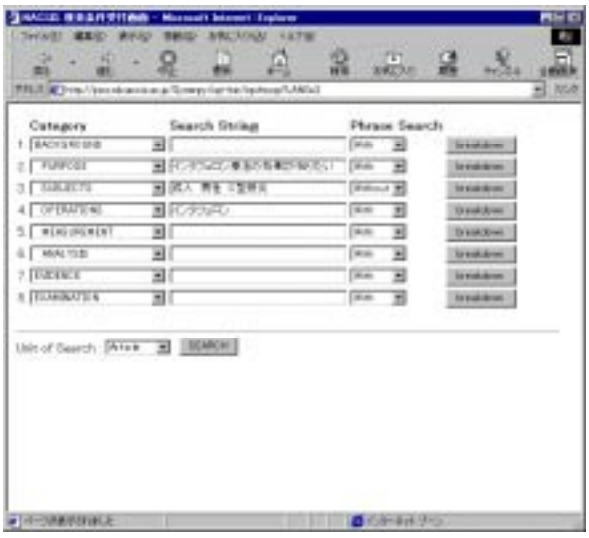
The ranked-output uses OpenText's "RankMode Relevance1", which ranks the members of the returned set based on the term frequency, the document length, and the total number of words. It is a variant of the standard *tf-idf*, which is used in most information retrieval systems.

4 Experiments

This section reports experiments on text retrieval of full-length texts or any passages, skimming and browsing in the retrieved document, and comparison of relevant passages (sentences) in the retrieved document across multiple documents using a structure-tagged full-text database of Japanese research papers. Although the system can process both Japanese and English documents and queries, Japanese documents and queries are used here since the workshop is entitled "Information Retrieval with Asian Languages".

4.1 Text Retrieval

The purpose of this experiment is to test the effectiveness of text-level structure as a role indicator for each concept in a query, by distinguishing the discourse types, and by identifying the passages suitable for synonym extraction.



ig. 3-1: Query Formulation with Structural Slot

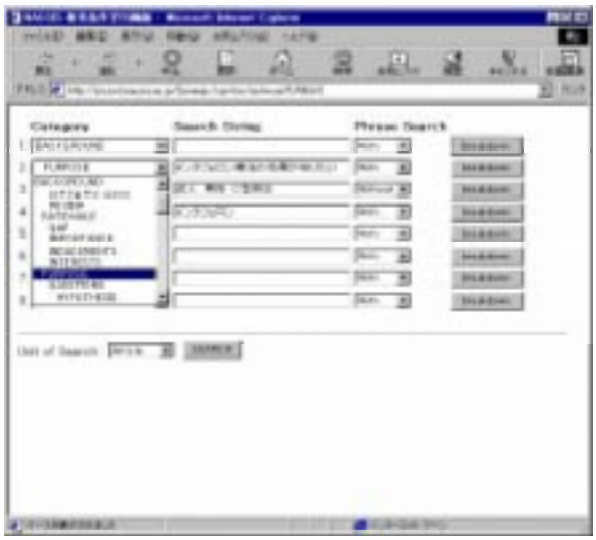


Fig. 3-2: Pull-down list of Category Labels

Fig. 3-3: Breakdown of input query

Fig. 3-4: Ranked List of Retrieved Documents

Fig. 3: Query Formulation

To assist query formulation, the system shows several slots for queries (Fig. 3-1)[2], based on a survey of user's search behavior (Kando, 1997c). Each slot has a category label. Natural language sentences can be entered in any slots and automatically analyzed into words and phrase, based on the method reported in (Kando *et al.*, 1998a). The default labels are second level categories, such as "A3 Research Topic", "B2 Subjects", "B4 Measurement", etc., and they can be customized. The categories can be changed by choosing another one from a pull down menu (Fig. 3-2). When no category labels are chosen, the system will apply default categories, the ones shown to perform best in (Kando, 1997a), to each term. When the "Breakdown" button is selected, the system provides the distribution of hits across categories for each input term so that the user can choose appropriate terms or categories based on them. (Fig. 3-3)

[2] (Kando, 1997c) compared the effectiveness of three models of query; (a) automatic, (b) interactive, and (c) structural slots as shown in Fig. 3-1. For (a), the user enters a natural language query and the system assigns category labels to each term. In (b), the system provides the distribution of hits across categories for each input search term, so the user can choose appropriate combination of terms and categories based on the distribution. The number of search terms entered by users was more than doubled with (c) than with an open entry box for (a) and (b). All the participants stated that the type (c) "structural slots" were helpful in recalling other important aspects of search topics and to examine and express their needs. This result agrees with Allen's studies, which showed that structural questions (reference question based on typical components of research papers) could derive more information from users than open questions or closed questions in the settings of reference interviews (Allen, 1988, 1989).

Search units, such as sentences, paragraphs, chapters or articles can be selected from a list box as well. In the ranking of the search results, extra weight is added to the term occurrence in the ***components*** specified in the query (Fig. 3-4).

4.2 Display Modes and Skimming

The documents selected from the ranked list can be displayed in various units, such as hit passage, paragraph (including the hit passage), chapter (including the hit passage), skimmed sentences, abstract (of the article that includes the hit passage) or the entire article regardless of the search unit of the text.

(1) SKIMMING

When "Skimming" is selected, the salient sentences are displayed in bright blue (dark tone in monochromatic print) and other sentences will be faded out in pale blue (pale tone in monochromatic print), so that the user can skim the text quickly to confirm the context around the salient part. The title of the article, chapter or section titles are also displayed (Fig. 4-1). Category labels can be added to the display (Fig. 4-2). Category labels are displayed in bright pink and bold face fonts.

Fig. 4-1: Skimming

Fig. 4-2: Skimming (with category label tags)[3]

(2) SKIMMED SENTENCES

When "Skimmed Sentences" are selected as the display unit, only the salient sentences will be displayed, forming a quasi-summary or text gist as a collection of the salient sentences (Fig. 4-3). Adding category labels (Fig. 4-4) or title and section titles to the gist (Fig. 4-5) or both will be helpful in understanding the text quickly.

[3] Colored version of figures will be available from author's homepage, http://www.rd.nacsis.ac.jp/~kando/

Fig.4-4 Skimmed Sentence with Category Labels

Fig.4-3 Skimmed Sentences as a Quasi-Summary.
Hit terms are displayed in red.

Fig.4-5 Skimmed Sentences with Section Titles

(3) BELIEF VALUE

A "skimming Belief Value" will be specified by the user. It is a threshold value. The sentences which contain the equal or higher belief values are displayed as skimmed sentences. A larger value will result in a shorter skimmed text gist.

Other available display options are "hit passage underlined" and "search term highlighting". If the search unit or the display unit is smaller than an article, the user can change the display mode into longer ones whenever desired. Through this operation, the user can verify the context of the hit passages or methods used in the study reported in the paper which includes the hit passages in order to examine the validity and reliability of the study.

4.3 Comparing Information across Multiple Texts

A search using text structure is more effective than ones that do not differentiate the role or function each concept plays in the text (Kando, 1997a). In particular, such searches are more effective for sentence or passage level retrieval. While passage retrieval is known to be more effective than ordinary full-text retrieval (Salton *et al.*, 1993), each passage is not an individual object. We often need to think of the passage's meaning in the larger context of the text as a whole. A passage, especially a sentence is too small to contain all the information being retrieved in the right context. Specifying the role or function each concept plays in the text, or context other than the retrieved sentences, is often useful to make the search more effective.

Moreover, such a small passage or sentence is short enough to read and understand instantly. Displaying the relevant sentences across multiple texts is therefore a convenient way to compare and analyze the differences or similarities among the texts for the matter in question. Figs. 5-1 and 5-2 show examples.

Fig. 5-1: Relevant sentences describing "Positive Ratio of HVC virus among Dialysis Patients"

Fig. 5-2: Relevant sentences answering the question "Are there difference of the effectiveness of Interferon therapy with age?" (On Japanese interface)

Fig. 5 Comparison among relevant sentences across multiple documents

Fig. 5-1 shows relevant sentences regarding the problem of "positive ratio of hepatitis type C virus in patients who have received artificial dialysis" and these are reported as the original evidence obtained by the studies reported in the articles (the category label is "C1 Evidence"). Using the categories, more specific retrieval is obtained. Otherwise, more noise sentences may be included in the retrieved passages. Also the list is short enough to compare all the sentences at a glance and the user can appreciate that "the positive ratio of HCV virus among dialysis patients is higher than that for ordinary people and is increasing". Again, the system enables the user to confirm the context of each passage, to skim the whole text or text gist, whenever desired.

Fig. 5-2 shows the answers to the question "Is there significant difference of effect of interferon therapy for viral hepatitis type C in adult males with age?" Such a complex search request cannot be expressed without structural slots or without structuring expressions: no single sentence contains every necessary concept in the request. In addition, with structural expression of the query, such high precision searches can be available.

In these examples, the tool supports the user's

information work or “reaggregation” of information scattered in the texts into one new document that allows comparison of the differences or similarities among the different texts, analysis of the trends among them, or compiling weak evidence from many texts to build strong evidence as a whole. These can be described as the discovery of unknown knowledge or relationships from the published texts (Swanson, 1986), or a kind of “text mining”. The system is especially appropriate for supporting such activity by users.

4.4 Recall/precision-based Evaluation

This section shows the results of searches of full-length texts using the text structure. Thirty-six queries are collected from researchers of the subject domain for the experiment purpose. Top 1000 documents are retrieved for each query in each method. The average over these thirty-six search queries are shown in Fig.6 and Table.1. The "Word" shows the retrieval results without text structure. The “Role” shows the retrieval results of which each term in the query is assigned appropriate category label or labels and “Dtype” shows those obtained by applying the default categories to every term in the query. These searches produce improvements in average precision of 32.3% and 31.8%, respectively, over searches without text structure.

These retrieval results suggests that not only is text structure effective in providing a flexible way of display or interaction, but it is also effective in traditional recall/precision based evaluation

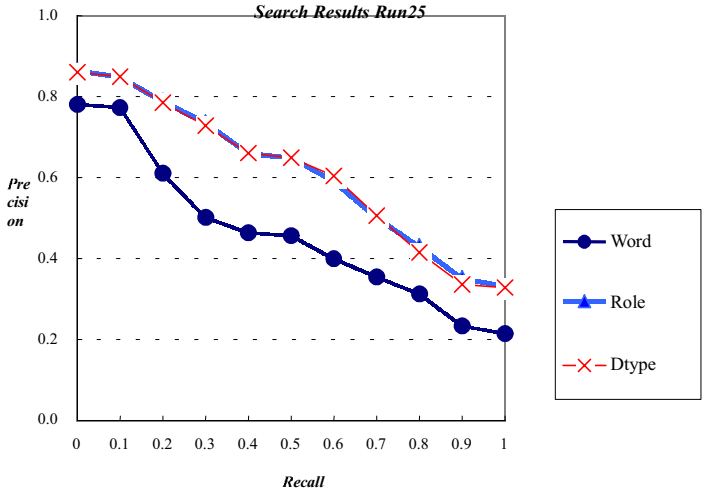


Figure 6. Search Results

Table 1. Effects of text structure (Role and Dtype) compared to searches not using it (Word) using 36 natural language queries

avarage of the search results of 36 natural language queries

	withoutStructure (Word,baseline)	Role	(%change)	Dtype	(%change)
0	0.7804	0.8629	(10.6%)	0.8616	(10.4%)
0.1	0.7732	0.8496	(9.9%)	0.8482	(9.7%)
0.2	0.6108	0.7870	(28.8%)	0.7857	(28.6%)
0.3	0.5004	0.7343	(46.7%)	0.7295	(45.8%)
0.4	0.4643	0.6616	(42.5%)	0.6618	(42.5%)
0.5	0.4570	0.6509	(42.4%)	0.6504	(42.3%)
0.6	0.3987	0.5909	(48.2%)	0.6020	(51.0%)
0.7	0.3531	0.5006	(41.8%)	0.5064	(43.4%)
0.8	0.3137	0.4299	(37.0%)	0.4163	(32.7%)
0.9	0.2351	0.3519	(49.7%)	0.3345	(42.3%)
1	0.2159	0.3338	(54.6%)	0.3266	(51.3%)
average*	0.4639	0.6139	(32.3%)	0.6112	(31.8%)

5 Future Study

So far the size of the experimental database used is small, with only a few hundred documents and limited types of text. Further investigation and evaluation are therefore needed to draw more accurate conclusions. However, the results reported in this paper seem sufficient to indicate the feasibility and implication of a retrieval system with rich post-retrieval process support functions, as well as a text

retrieval system with effective search in the traditional recall/ precision-based evaluation. For further investigation, the next steps are retrieval experiments using a large scale collection with real users. How to evaluate the effectiveness of these post retrieval processes is a challenging issue.

Acknowledgment

This research is supported by "Research for the Future" Program JSPS-RFTF96P00602 of the Japan Society for the Promotion of Science.

References

Allen, B. (1988) Text structures and the user-intermediary interaction. RQ, Vol.27, No.4, p.535–541.

Allen, B. (1989) Recall cues in known-item retrieval. Journal of the American Society for Information Science, Vol.40, No.4, p.246–252.

Biber, D. (1994) "13. Intra-textual variation within medical research articles" Corpus-based Research into Language. Edited by Oostdijk. Rodoph, Al Lanta, p. 201–221

Bishop, A. P. (1998) "Digital libraries and knowledge disaggregation: the use of journal article components", In. Proceeding of ACM Digital Libraries '98. p. 29–39.

Bishop, A. P. (1999) Document structure and digital libraries: how researchers mobilize information in journal articles. Information Processing and Management, Vol.35, No.3, p. 255–282.

Day, R. A. (1988) How to Write and Publish a Scientific Paper, 3rd ed. Oryx Press, 211 p.

Dillon, A. (1994) Designing Usable Electronic Text: Ergonomic Aspects of Human Information Usage. Taylor and Francis. 195 p.

Endres-Neggemeyer, B., Maier, E. and Sigel, A. (1995) How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor. Information Processing and Management, Vol.31, No.5, p.631–374.

Fuller, S. S. (1984) Schema Theory in the Representation and Analysis of Text. Ph.D. thesis, Univ. Southern California, 189 p. available from U.M.I. Order No.DA8500206.

Jang, D. H., and Myaeng, S. H. (1997) "Development of a document Summarization System for Effective Information Services", In. Proceeding of RIAO '97, Montreal, Canada.

Jones, P. A. (1995) Automatic Abstracting and Indexing of Technical Documents: an Approach Based on Concept Selection. Ph.D. dissertation, Lancaster University.

Kando, N (1991) Structure Analysis of Information Media using Categories: Structure of Research Articles. Tokyo, Keio University. Master's thesis, 227 p. (in Japanese).

Kando, N. (1995a) Text Structure of Information Media: as a Framework for Content Analysis. Tokyo, Keio University. Ph.D. Thesis, 256 p. (in Japanese).

Kando, N. (1995b) Structure of news stories: as relating to indexing and retrieval. Journal of Japan Indexers Association. Vol.19, No.1, p.1–17 (in Japanese).

Kando, N. (1996) Text structure analysis based on human recognition: cases of Japanese newspaper articles and English newspaper articles. Bulletin of the National Center for Science Information Systems. No. 8, p.107–129 (in Japanese, with English abstract).

Kando, N. (1997a) "Text-level structure of research articles and its implication for text-based information processing systems". In Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research, Aberdeen, Scotland, UK. p.68-81.

Kando, N. (1997b) "Text-level structure: Implications for Information Retrieval and the Potential for Genre Analysis". Paper presented at the NLP Seminar, Sheffield University, Sheffield, UK, 24 p.

Kando, N. (1997c) "An approach for text information retrieval, browsing, and extraction using discourse-level structure", [Poster] Paper presented in the 20th annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA.

Kando, N., Kageura, K., Yoshioka, M. and Oyama, K. (1998a) Phrase processing methods for Japanese text retrieval. SIGIR Forum, Vol.32, No.2, p.23-28.

Lehman, A. (1999) Text structuration leading to an automatic summary. Information Processing and Management, Vol.35, No.2, p.181–191.

Liddy, E. D. (1991) The discourse level structure of empirical abstracts; an Exploratory Study. Information Processing and Management, Vol.27, No.1, p.55–81.

OpenText (1997) Livelink Index Engine Query Language Reference. OpenText Corporation.

Miike, S, Itoh, E, Ono, K and Sumita, K. (1994) "A full-text retrieval system with a dynamic abstract generation function", In. Proceedings of the 17th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, p.152-161.

.Moens, M. F., and Uyttendaele, C. (1997) Automatic text structuring and categorization as a first step in summarizing legal cases. Information Processing and Management, Vol.33, No.6, p.727–737.

Oddy, R. N., Liddy, E. D, Balakrichnan B, Bishop, A, Elewononi, J. and Martin, E. (1992) Towards the use of situational information in information retrieval. Journal of Documentation, Vol.48, No.2, p.123–171.

Oakes, M. P. and Paice, C. D. (1999) "The automatic generation of templates for automatic abstracting" In. Proceedings of the 21st Annual Colloquium of the BCS-IRSG, Glasgow, UK.

Paice, C. D. and Jones, P. (1993) "The Identification of important concepts in highly structured technical papers". In. Proceeding of the 16th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA., p.69-78.

Rama, D. V. and Srinivasan, P. (1993) An investigation of content representation using text grammars. ACM Transactions on Information Systems. 11(1):51–75.

Rau, L. F. and Jabobs, P. (1990) SCISOR: Extracting information from on-line news. Communications of the ACM. Vol.33, No.1, p.88–97.

Salton, G., Allan, J. and Buckley, C. (1993) "Approaches to passage retrieval in full text information systems". In. Proceeding of the 16th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA., p.49-58.

Samuels, S. J., Tennyson, R., Sax, L., Mulcahy, P., Schermer, N. and Hayovy, H. . (1988) Adults' use of text structure in the recall of a scientific journal article. Journal of Education Research. Vol.18, No.3,171–174.

Swanson, D. R. (1986) Undiscovered public knowledge. Library Quarterly. Vol.56, No.2, p.103–118.

Strzalkowski, T. (ed.) (1999) Natural Language Information Retrieval. Kluwer Academic Publishers, 384 p.

Swales, J. M. (1990) Genre Analysis: English in academic and research settings. Cambridge, Cambridge UP, 260 p.

Teufel, S. and Moens, M. (1997) "Sentence extraction as a classification task". In Proceedings of the Workshop on Intelligent Scalable Text Summarization, in association with ACL/EACL-97.

Ueda, S, Koshizuka, M. and Kando, N. (1995) Framework for image recognition and alternative indexing method for image. IPSJ SIG Notes (95-CH-28). Vol.95, No.96, p.55–60 (in Japanese, with English abstract).

van Dijk, T. A. (1980) Macrostructures: an Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition. Hillsdale, Lawrence Erlbaum Assoc., 317 p.