Penstation: Easy Access to Relevant Facts without Retrieving

Kazunori Muraki and Kenji Satoh
Information Technology Research Laboratories
NEC Corporation
1-1, Miyazaki 4-chome, Miyamae-ku, Kawasaki, Kanagawa, 216, Japan
k-muraki@hum.cl.nec.co.jp, satoh@hum.cl.nec.co.jp

ABSTRACT

In this paper, we propose a document-preparation environment integrated with text-retrieval in the frame work of seamless digital network library. We often write documents on a word processor while referring to other text-files on personal files , group files and much open information resources like libraries and information serviced on digital network as well as books, papers and printed matters. Much of information has been already on disks and networks so that tedious looking-up work of books and paper dictionary is largely reduced with mouse clicks or key strokes for retrieving digital information resources . But there are still difficulties in referring along with writing documents on computer: (1)The acts such as key operations, mouse operations for file retrieval and(/or) book consultation for detecting relevant references interrupt users' concentration on how to express what they have in their heads, (2)No practical way to find location of information is available although users know surely there exists a vast volume of indispensable information for explicating their ideas somewhere, (3)Unconsciousness of existence of useful information can not invoke users' request of information retrieval. To solve these problems, we propose an integrated system to realize the automatic and low-cost index acquisition and documentation support environment by utilizing rich indexes acquired during writing documents. This environment encourages users' building ideas by low-cost presentation of document components collected from personal, group and open information resources. A sophisticated idea building environment has been realized based on the concept of episodic indexes and citation indexes for consulting easily personal and group document bases. It is also supported by functional retrieving mechanism of much larger volume of text data base. These establishes a new environment integrated with episode indexing which enables the total utilization of both known in personal and group files, and unknown and new information in open data base.

KEYWORDS: Information Retrieval, Citation, History, Index

INTRODUCTION

We have many chances today to prepare documents with a personal computer or on a word processor. There are a variety of documentation support tools such as outline processors[2], idea processor [3, 1] and groupwares[13, 7]. Recently, there has been research for computer supported creation[4, 8, 14] as well. But these systems are not for expressing ideas on texts , not for developing ideas through writing lines of texts but only for handling words or formats of texts. They can neither process sentences or ideas themselves, nor improve or develop them.

We express our ideas concretely using sentences when we have already in mind what to express. Furthermore, we grow our ideas while we are thinking of what to write. Therefore support facilities for a computer user to prepare documents help him/her to build up ideas so much if they are suitable and natural to expressing ideas.

A convenient environment to document writers has been gradually being created where we can, freely and easily, refer to a variety of information, for example, an encyclopedia on the terminal of a computer, which is of much help for document preparation and idea building. However reference of information along with writing documents causes the following deficiencies:

- The key or mouse operations as well as key-word inputting to invoke retrieving texts interrupt concentration on what to write.
- No easy access to information on computers of others is available although user knows clearly that information in their discs is indispensable for efficient group work.
- Unconsciousness of existence of useful and relevant information can not invoke users' making request of retrieval.

Behind them are the facts [5]:

1. Key strokes in writing sentences for expressing an idea are far different from key strokes for expressing keywords for retrieval, so that keyword input for retrieving disturbs a way of thinking.

2. No support facility is available for people to use information resources developed and recorded by others in a group.

3. Explicit retrieval action by inputting keywords is imperative.

This paper proposes an integrated system to realize an automatic episod index acquisition and an easy information sharing capability based on the episodic and citation based information retrieval mechanism as well as an automatic document retrieval capability with functional descriptions as keywords for retrieving relevant facts. [11, 16].

We propose the following three fundamental concepts which provide three way of accessing relevant information more easily.

Personal Episodic Information in your Desk :

Non content indexing scheme derived from any kind of action ,episode as we call it here, to provide discriminative access key to relevant text even among texts of similar content.

In-group Information in your Book Shelves :

Revese citation which records citation of a work and reports the citation to the original writer is installed to promote information sharing within a group.

Open Information Space in Library :

Automatic retrieval capability of useful information in parallel with writing text is defined on functional description.

We named this environment "Penstation", because this environment is a station which calls together and dispatches information.

EPISODE INDEX

We often don't know where to search for the text that even we ourselves have written or read. In this case, we look for the title of the chapter or article in our obscure memory using a text editor [12], or we execute a full-text search[15]. In many occasions, however, we often remember the fact of searching for a text by some key word which serves as a clue, or the fact of cutting and pasting some part of it , instead of the very contents of the text we need. These memories, that is to say, the episodes regarding the text provide easier and quicker retrieval of it since they are clearer in our memories than the detailed description of the text [9].

To minimize the extra cost for indexing the episodes, it is necessary to extract the episodes out of the history of text preparation and reference and to have the text hold a large amount of index of them. This non-content indexing co-exists with content indexing.

A general model of this indexing in the application is illustrated in Figure 1. This model is designed for the indexing module to supervise the user's operation and the text throughout the processing of the application from inputting, editing to outputting, so that it extracts the history of all the operations and gives the index of the episodes to the text concerned.

The seven basic operations and the extracted item of the history by each of the operations are shown in the Table 1. Figure 2 indicates the actual operations to create those basic operations.

The function of episodic retrieval is available in the application outfitted with the Episode Index Acquisition Model. Figure 3 is an example index acquired by this model in a printout application.

REVERSE EXTRACTION OF CITATION

One of the episodes in the previous section, the fact of being referred to or quoted is important to the author of the text. When preparing documents concerning research, patents, business, we search for any existing works whose contents are related to ours. If there are any similarities, we try to distinguish our own ideas from that of others. In addition, citation history of one's own works is equivalent to attention paid by others, whether it be a criticism, evaluation, interpretation or approval.

The model in the previous section is usable for extracting citation history if the following problems are solved.

- While the episode index is designed for personal use, it is impractical for an individual to control such a vast amount of data with regard to academic reports and patent rights.
- Since this extraction doesn't always happen simultaneously with the author's computer operations, indexing the text is useless for transferring the information.

The first problem is resolved with integrated text bases. All the computers are to be supplied with this index acquisition model as the server-client roles are dynamically exchangeable, even for a distributed system.

The second problem is easy to settle if the environment is arranged for distributed index acquisition. When the information of being cited is sent to the machine to process the text, that information is transferred to the computer from which the text was originally edited, and consequently to the author of the text. This notification can be received promptly as an electronic mail arrival sign, or instantly at the time of the user's monitoring. The notification mode is to be customized according to the categories of the information of being cited: mere reference or quotation, criticism or other comments.

An important point to remember is that any evaluations of cited text is designed to broaden the knowledge of a person who cites the text; regardless of the cost or benefit of the author of the text. In the distributed index acquisition environment, all the participants who open their own works to others can develop their ideas widely with the help of others' various comments.

An example is illustrated in Figure 4 in which Mr. A cites to my Text 0 while making Text 1. Mr. A's citation corresponds to "1 : Reverse Citation Extraction", and "2 : Generation of Reverse Citation Index" is executed on my text, and "3 : Reverse Citation Notification" notifies me instantly of my Text 0 being cited by Mr. A in Text1.

AUTOMATIC RETRIEVAL OF RELATED INFORMATION

As stated in the previous section, we usually prepare research papers or patent documents, being fully aware of the preceding cases concerned. We in general search for the necessary information depending a key word. We often fail to find closely related information since such information is not always searched for by a key word in our mind and our documents change when our ideas grow more sophisticated. A completed sentence which has a logical meaning may has crucial information of key words more suitable to relevant information. A newly inputted word sequence in a context may serve for advantageous retrieval keys as stated below:

1. Words combined by function words like prepositions(case markers), and predicates(verbs and adjectives) appearing in texts may have meaningful filtering function of relevant facts than the simple logical combination of keywords for text retrieval.
2. Functional description which is composed of two words combined by a functional word(a case maker), or is composed of two words combined by a predicate is automatically extracted easily and at a high quality.

We constructed an environment, which retrieves related documents synchronized with text inputting, as shown below in Figure 5.

Retrieval Element Extraction

In conventional text retrieval systems, words, phrases, case markers, clauses or sentences could possibly be used for keywords. Words and phrases without context has a restricted effect to reduce the number of irrelevant texts. Clauses and sentences as keywords on the contrary rarely makes matching on the surface between the keywords and relevant texts. Japanese fundamental sentence is composed of a predicate and its phrases with function word, each of which is called a case element. We utilize the logical function of case elements in automatic related fact retrieval, instead of words and clauses[10].

Real-time and high-efficiency parsing is beyond the latest NLP technology. But as a Japanese case element as a rule appears on the surface before a case marker, a pattern of "noun + a case marker ... predicate(verb/adjective)" is correctly extracted form the input sentences to the degree needed.

For extracting retrieval keywords using this pattern, we employed an well known Japanese morphological parsing technology [6] in order to comprehend parts of phrase and partition between words. This parsing extracts easily a retrieval keyword as a pattern by receiving parts of phrasal structures.

At the moment of finishing inputting 「自民党に離党届を提出した議員は、」(" 自民党 "= the Liberal-Democrats, " 離党届 "= a notice of party resignation, " 提出した "= sent (in a notice), " 議員 "= a Diet member, 「Diet members who sent in a notice of party resignation to the Liberal-Democrats」), two retrieval keywords are extracted as follows: 「自民党—に—提出」(「the Liberal-Democrats—に (target)— sent」) and「離党届—を—提出」(「a notice of leave the party—を (object)— sent」).

Display of Information

In a retrieving process, if the retrieval keywords found, different conjugation of predicates and incomplete matched nouns are also retrieved. One retrieval keyword "hits" two or more parts in the same text with a sentence as a unit. The advantage of a sentence as a unit to be displayed is:

- Easy to specify and display the parts to be searched
- Able to transfer the correct information.

In the present system, one retrieved sentence is displayed on one window, up to the maximum size of the screen. In case many sentences are searched, the priority calculation gives an order according to the complete matching of the noun and the conjugated form of the predicate. And a group of sentences retrieved by the same element are displayed on the same windows. The ones over the space of the window are unseen but can displayed by at the user's command.

The priority calculation grants priority and displays the sentence which corresponds to more than one retrieval keywords accordingly. The two retrieval keywords in the examples mentioned above, correspond to the following sentences in the source.

「自民党に椎名素夫参院議員は二十三日午前、離党届を提出した。」
(Mr. Shiina Motoo who is a member of the House of Councilors, sent in a notice of party resignation on the morning of 23th.)
「武村正義代議士ら１１人が離党届を提出し、」
(Eleven members of the House of Representatives including Mr. Takemura send in a notice of party resignation,)
「羽田派衆議院議員全員を含む４４人の離党届を梶山幹事長に提出した。」
(Forty-four members of the House of Representatives, including all the members of Mr. Hada faction, sent in a notice of party resignation to chief secretary Kajiyama.)

When matching of conjugation is considered, the first and the third examples above are prior to the second. Moreover the first, which includes both elements, is prior to the third. See Figure 6.

CONCLUDING REMARKS

We have introduced the automatic index acquisition/document preparation support environment to encourage users' building ideas by means of supporting much easier and cheaper information retrieval and sharing. We establish a new environment integrated with episode indexing, reverse citation, automatic acquisition of retrieval keywords, which enables the total utilization of both known and unknown information in three levels of information acquisition; 1)Personal document DB by episodic retrieval, 2)Group work DB by information sharing cultivated on reverse citation information flow management, and 3)Widely opened information base by automatic relevant fact retrieval utilizing functional descriptions specifying relevant information more directly.

REFERENCES

1. Chandrasekaran, B. 1986. Generic Tasks in Knowledge-Based Reasoning: High-Level Building blocks for expert system design. *IEEE Expert*, Fall, pages 23-30
2. Carol O. Day, etc. 1989. The Write Stuff PC/COMPUTING, pages 95-112, Dec.

3. William Hershey. 1985. Idea Processors BYTE, pages 337-350, June.

4. Susumu Kunifuji. 1993. A Survey on Creative Thinking Support Systems and the Issues for Developing Them. *Journal of JSAI* Vol. 8, No. 5, Sep.

5. Kazunori Muraki. 1992. Access Methods for Large Text-Base. *Proceedings of the 7th Annual Conference of JSAI* 7-8.

6. Naoyuki Nomura and Kazunori Muraki. 1991. A Documentation Aid Using Retrieval and Processing Units of Stratificational Linguistic Structures. *IPSJ SIG Notes* 91-NL-82-10.

7. J. F. Nunamaker, Alan R. Dennis, Joseph S. Valacich, Douglas R. Vogel and Joey F. George. 1991. Electronic Meeting Systems to Support Group Work *Comm. ACM* Vol. 34, No. 7 July.

8. Ryohei Orihara 1993. Trends of Systems Supporting Divergent Thinking. *Journal of JSAI* Vol. 8, No. 5, Sep.

9. Kenji Satoh and Kazunori Muraki. 1992. Personal Text Base. *Proceedings of the 6th Annual Conference of JSAI* 7-9.

10. Kenji Satoh, Akitoshi Okumura and Kazunori Muraki. 1992. Self-Navigational Access for Very Large Text Data. *Proceedings of the 45th Annual Conference of IPSJ* 5B-8.

11. Kenji Satoh and Kazunori Muraki. 1993. Detection of Relevant Facts without Retrieving. *Proceedings of the 7th Annual Conference of JSAI* 11-5.

12. Akio Shibata, Toru Kubo and Tsutomu Hata. 1992. Functional Specifications for an ODA Converter System *Proceedings of the 45th Annual Conference of IPSJ* 5B-2.

13. Mark Stefik, Gregg Foster, Daniel G. Bobrow, Kenneth Kahn, Stan Lanning and Lucy Suchman. 1987. Beyond The Chalkboard: Computer Support For Collaboration and Problem Solving in Meetings. *Comm. ACM* Vol. 30, No. 1, Jan.

14. Kozo Sugiyama. 1993. R & D Trends of Tools for Convergent Thinking: Computer Support Based on KJ Method. *Journal of JSAI* Vol. 8, No. 5, Sep.

15. Kousuke Takahashi, Hachiro Yamada and Masato Motomura. 1989. On Text Search Hardwares for Full-Text Database Systems. *IPSJ SIG Notes* 89-FI-14-5

16. Kenji Satoh and Kazunori Muraki. 1993. Penstation for Idea Processing. *Proceedings of NLPRS '93* pages 153-158.

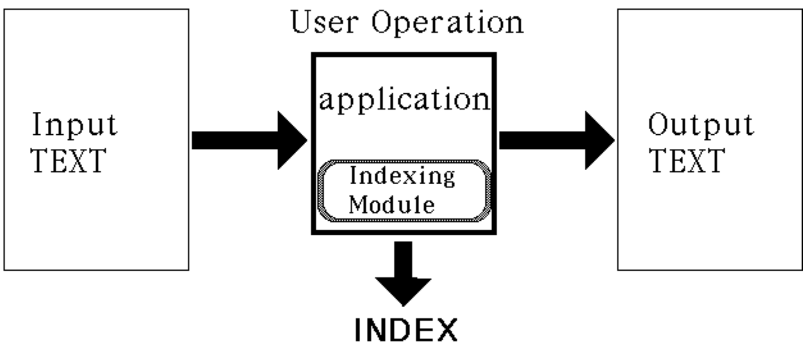

Figure 1: Episode Index Acquisition Model

Search(Get) Text	Time, Terminal, Operator, Search Key(File Name)
Citation	Time, Terminal, Operator, Referred Text Name
Reverse Citation	Time, Terminal, Operator, Used Text Name
String Replacement	Time, Terminal, Operator, Replace & With String
String Search	Time, Terminal, Operator, Search String
String Cut/Copy/Paste	Time, Terminal, Operator, String
Save/Print	Time, Terminal, Operator, File/Printer Name

Table 1: User's Operation Extracted as Episode

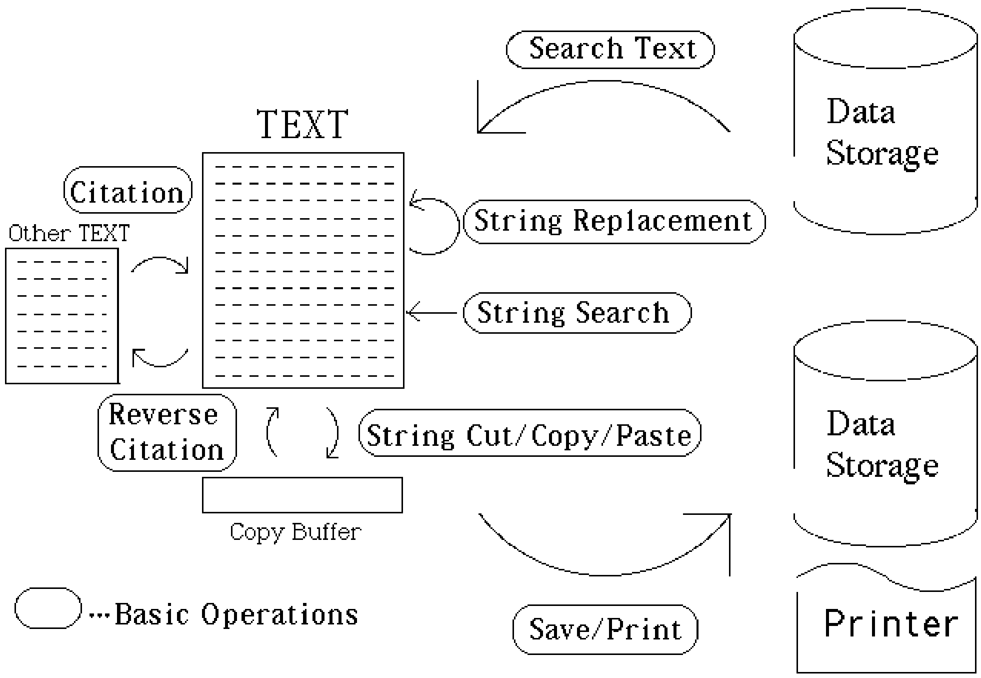

Figure 2: Basic Operations in Episode Index Acquisition Model

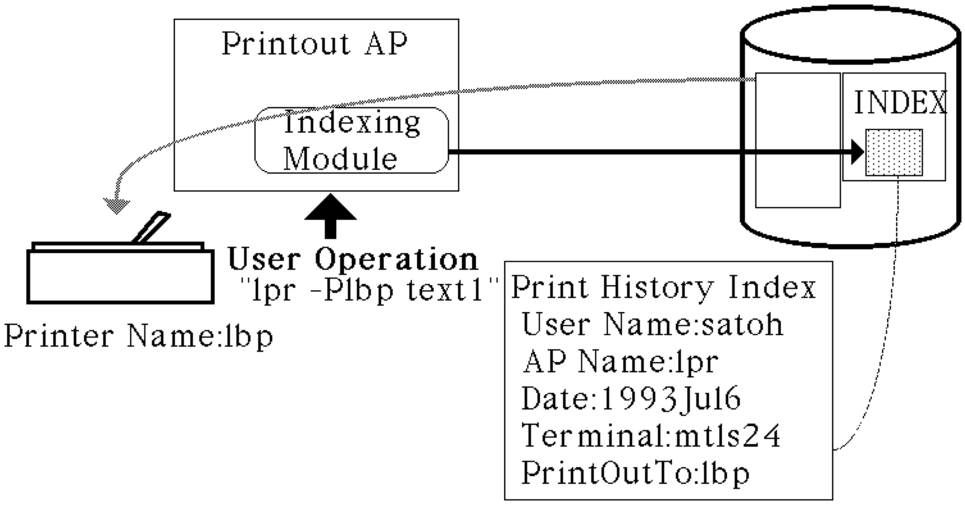

Figure 3: The Model Supplied in Printout Application

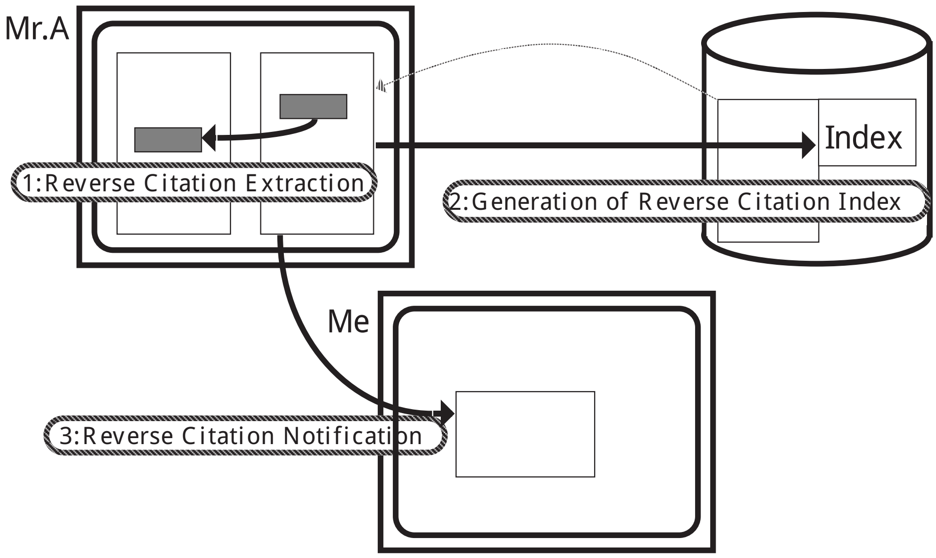

Figure 4: Processing of Reverse Citation Information

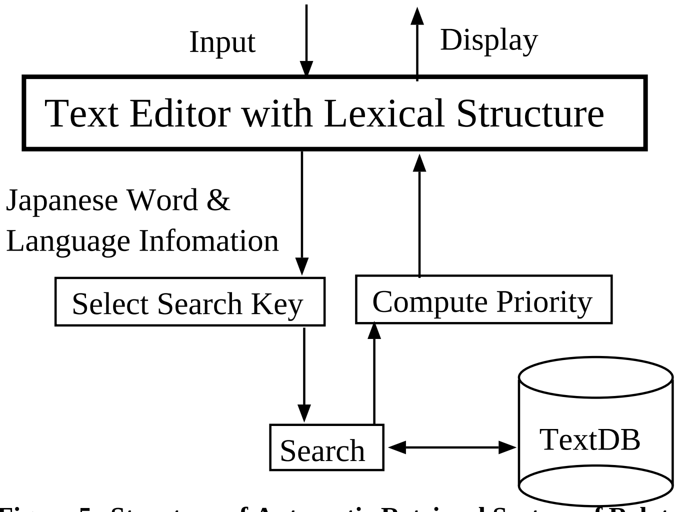

Figure 5: Structure of Automatic Retrieval System of Related Information

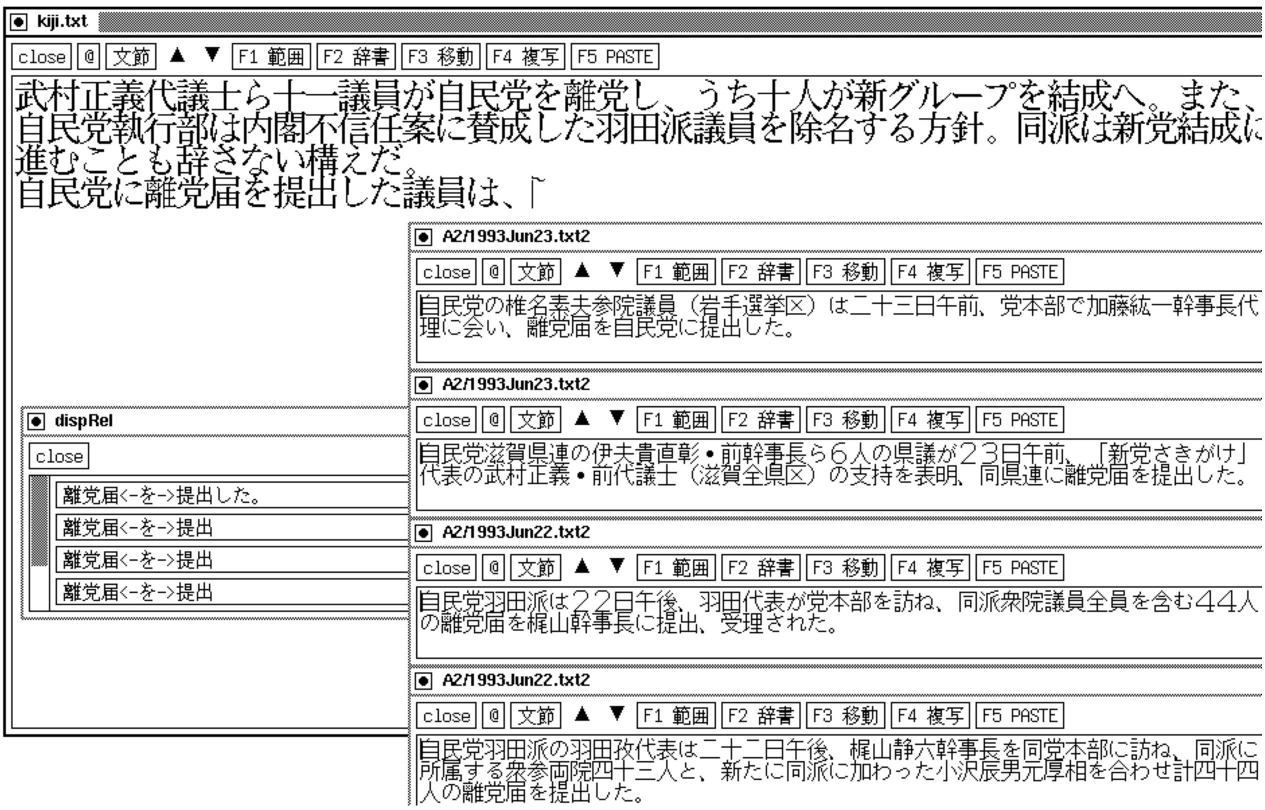

Figure 6: System Operations