Construction of a Dynamic Thesaurus and Its Use for Associated Information Retrieval

Haruo Kimoto Toshiaki Iwadera

Nippon Telegraph and Telephone Corporation
Electrical Communication Laboratories
407C 1-2356 Take, Yokosuka-Shi
Kanagawa 238-03 Japan

Electronic mail: kimoto%nttnly.ntt.jp@relay.cs.net
Electronic mail: iwadera%nttnly.ntt.jp@relay.cs.net

Abstract

An information retrieval system based on a dynamic thesaurus was developed utilizing the connectionist approach. The dynamic thesaurus consists of nodes, which represent each term of a thesaurus, and links, which represent the connections between nodes. Term information that is automatically extracted from user's relevant documents is used to change node weights and generate links. Node weights and links reflect a user's particular interest. A document retrieval experiment was conducted in which both a high recall rate and a high precision rate were achieved.

The topics discussed in this paper:

Connectionist Model, Automatic Indexing, Information Retrieval, and Thesaurus.

1. Introduction

The development of word-processors, optical disk filing systems, and computer networks enables us to use large scale databases. In order to advance database systems, there are two problems that need to be addressed: document storing and document retrieval. For document storing, an automatic document classification system and an automatic indexing system have already been developed.[1][2] For document retrieval, many kinds of AI techniques are currently being studied. The AI techniques most commonly used in a database system are expert systems[3] and natural language processing systems.[4] Development of an expert system requires a rule base for each application. Natural language processing techniques have difficulties with syntactic analysis and semantic analysis.

In this paper, a new system is proposed which incorporates the connectionist model in a dynamic thesaurus. This system is designed to discover and use the interests of a user so that the results of document retrieval are more beneficial to that user. This new system is called the Associated Information Retrieval System (AIRS).

2. Basic Concept of the AIRS System

The general diagram of AIRS is shown in Fig.1. The distinctive feature of AIRS is that it determines the user's interests from user's sample relevant documents as term information. This term information is used to construct a dynamic thesaurus that generates associated keywords. These keywords are used to retrieve documents that precisely fit the user's own interest.

2.1 Term Information from User's Sample Relevant Documents

Term information consists of keyword information and keyword relation information. Keyword information consists of keywords and the ranking of each keyword, which is ranked according to the importance of that keyword in a particular sample relevant document. Keyword relation information consists of relation type and relation strength. These are shown in Fig.2.

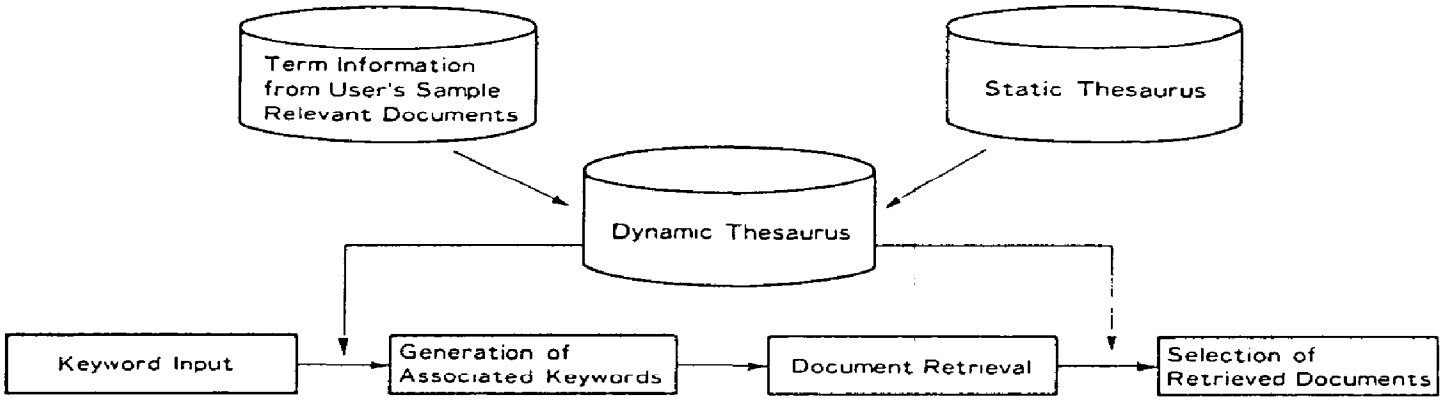

Fig. 1 The General Diagram of the Associated Information Retrieval System (AIRS)

2.2 The Dynamic Thesaurus

The dynamic thesaurus is constructed based on a network structure. Each node of the network, which has a node weight, represents one term of the thesaurus. Each link represents a relationship between terms. The data structure of the dynamic thesaurus is shown in Fig.2. Nodes (Term and Node Weight) and Links (Relation and Link weight), which constitute the dynamic thesaurus, reflect a user's interest. The node weight of a term is calculated using the keyword ranking in term information. There are five kinds of relations between nodes. They are as follows:

a. Broader term relation
b. Narrower term relation
c. Use relation(Descriptor)
d. Used for relation(Synonym)
e. Co-occurrence relation

Relations of a, b, c, and d are obtained from the static thesaurus. Relation e, the co-occurrence relation, is obtained from keyword relations in term information. The co-occurrence relation is defined as the relation of keyword pairs that co-occur in the sample documents (See Fig.3).

There are a lot of small, separate networks in the initial state of the dynamic thesaurus, which, initially, is identical to the static thesaurus. The use of co-occurrence relations in the dynamic thesaurus makes it possible to personalize the thesaurus by modifying the node weights and links. AIRS uses the dynamic thesaurus to generate associated keywords from the input keywords of a user.

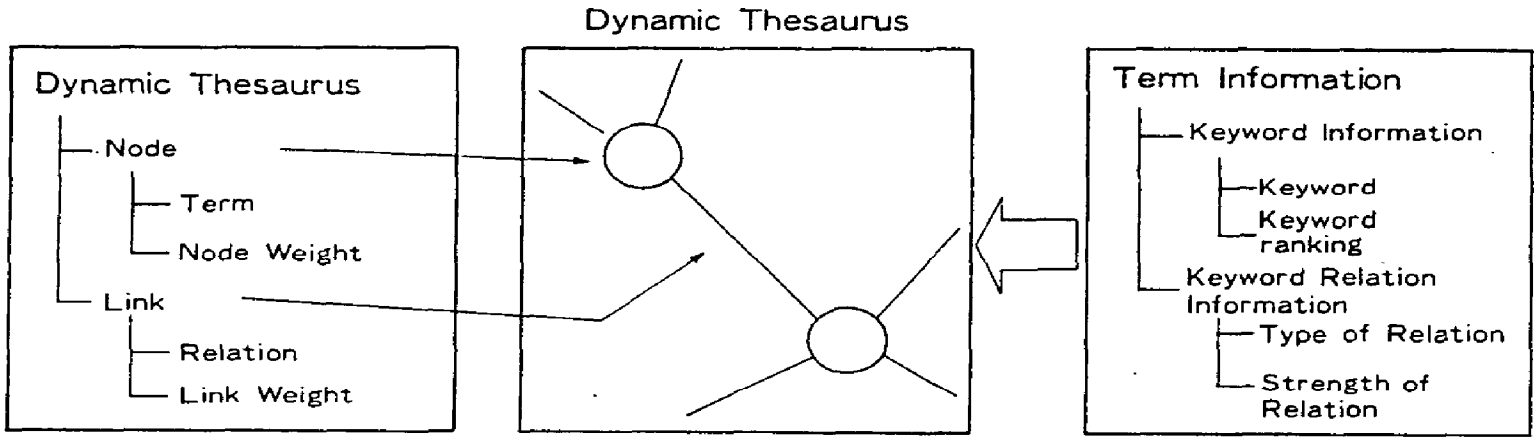

Fig. 2 Construction of The Dynamic Thesaurus using Term Information

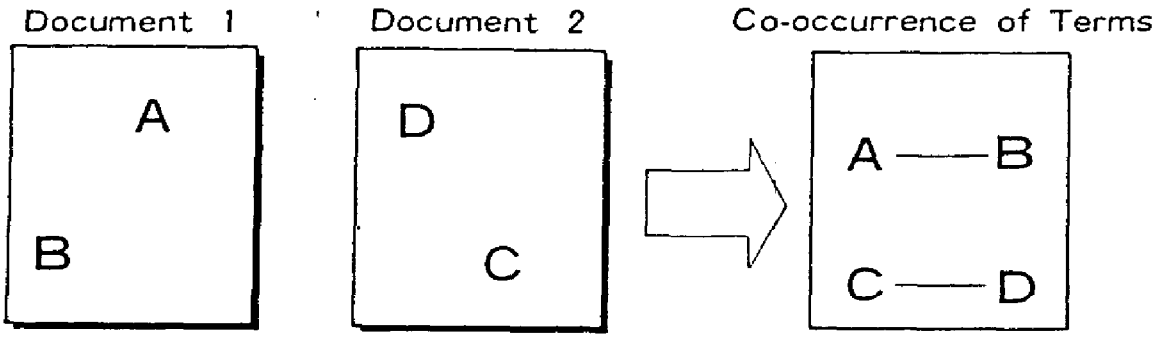

Fig. 3 Co-occurrence of Keywords in Documents

2.3 Expected Effect of the AIRS System

The expected effects of AIRS are as follows:

(1) Associated keywords, which reflect the user's interest, are generated by the dynamic thesaurus.

(2) Both a high recall rate and a high precision rate are achieved by using these associated keywords for document retrieval.

(3) It is possible to use state transitions of the dynamic thesaurus to reflect a user's change of interest over time. Thus, document retrieval reflecting the user's prior interests is possible.

3. An Overview of the AIRS System

The general operating procedure of AIRS is described in this section. The configuration of the prototype system is shown in Fig.4. The system operates as follows.

Step 1. Term information is extracted au tomatically from the user's sample documents. Keywords, keyword ranking, and keyword co-occurrence information are extracted using the INDEXER system.[2]

Step 2. A static or traditional thesaurus is modified by term information to form the dynamic thesaurus. Links are generated and node weights are calculated while the dynamic thesaurus is being made.

Step 3. Associated keywords are generated from a user's input keyword using the dynamic thesaurus. The dynamic thesaurus starts with an input keyword and then selects associated keywords based on their node weights and links.

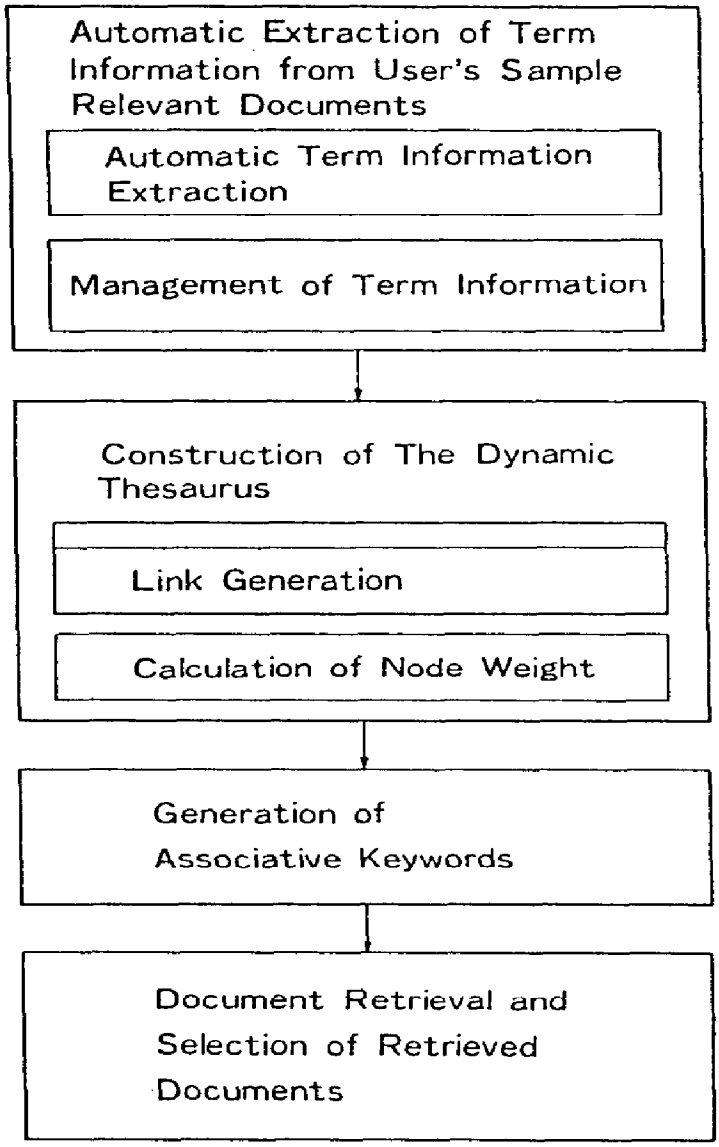

Fig. 4 AIRS Operating Procedure

Step 4. Documents are retrieved using these associated keywords. Retrieved documents are ranked by using information in the dynamic thesaurus.

The following section describes each of these functions in detail.

4. Algorithms

4.1 Link Generation Algorithm

If two keywords occur in a document, links are generated between corresponding nodes in the dynamic thesaurus if no previous link exists between these two nodes (See Fig.5).

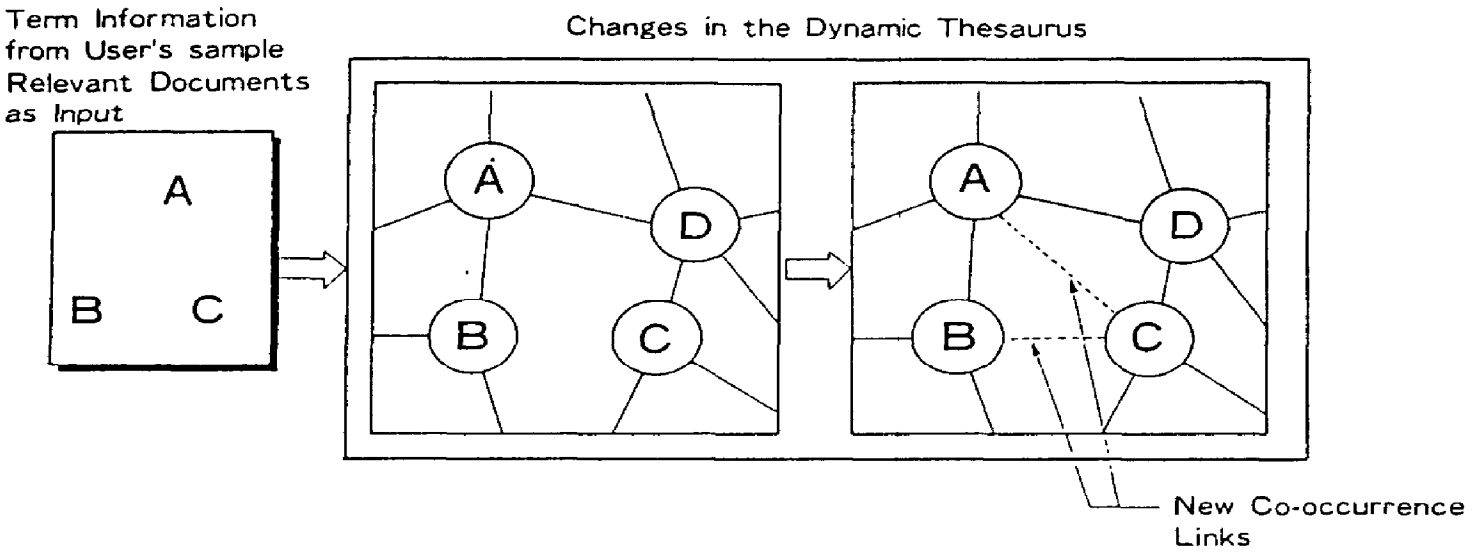

Fig. 5 Generation of Links

4.2 Node Weight Calculating Algorithm

Node weight reflects the importance of the keywords extracted from the user's sample documents. The importance of the keywords is calculated by the INDEXER system. The INDEXER system extracts and ranks keywords from each of the user's sample documents. Ranking is in the order of importance according to frequency and location in that document. In the formulas presented below, (1)-(4), D is the complete set of sample documents, and Ti denotes individual documents. Ki is the set of keywords extracted from Ti with the INDEXER system. KWij denotes individual keywords. Assume there are n documents and a total of m keywords in a document Ti; hence,

$$D = \{Ti\} \quad (i=1,\ldots,n) \text{ and} \qquad (1)$$

$$Ki = \{KWij\} \quad (j=1,\ldots,m). \qquad (2)$$

The importance of KWij to Ti is denoted as KI(ij). The value of KI(ij) is designed to decrease linearly as J, the ranking number, increases, and the sum of the value of KI(ij) (j=1 to m) equals one (for each i) for normalizing the importance; the closer the value of the keyword is to one, the more important the keyword is ranked. KI(ij) is calculated using formula (3):

$$KI(ij) = \frac{2TWi}{m*(m+1)} * (m+1-j), \qquad (3)$$

where TWi is the value given to Ti in D, and j is the ranking of KWij in Ti. TWi is calculated using the following formula:

$$TWi = \frac{DW}{n}, \qquad \text{(Constant)} \qquad (4)$$

where DW is the value of D given by a user.
After KI(ij) is calculated for each Ti, the node weight of node l, denoted as NWl, is calculated as the sum of KI(ij) in D.

4.3 Associated Keyword Generation Algorithm

Associated keywords are intended to extend the keywords inputted by the user. Associated keywords are obtained by traversing the links and nodes of the dynamic thesaurus starting with the node that corresponds to the user inputted keyword. Hereafter, in this paper, the starting node is called the "generation starting node," and the set of links and nodes traversed in the associated keyword generation process is called the "generation path." The traversing distance is defined as the number of links traversed to generate an associated keyword. The AIRS procedure for associated keyword generation is as follows:

Step 1: The traversing distance and the kinds of links to traverse are preset. The threshold value of the node weight is preset for selecting nodes, which are likely to be generated as associated keywords. The distance between two nodes in the dynamic thesaurus is defined as the number of links between those two nodes.

Step 2: Starting from the generation starting node, acceptable links are traversed up to the preset distance.

Step 3: All nodes in the generation path become candidates of associated keywords.

Step 4: Among the candidate nodes, only those that have a node weight larger than the threshold value are outputted as associated keywords.

An example of the keyword generation process is given in Fig.6. Assume that the traversing distance is set at three, that all kinds of links can be traversed, and that the traversing is limited to the enclosed area in Fig.6. The generation starting node is node A. The generation path consists of nodes A, B, C, and D, which also become candidate nodes. Finally, nodes A and D, whose node weights are larger than the threshold value, are selected as associated keywords.

5. Experimental Results

A prototype of AIRS was created. The configuration of the prototype system is shown in Fig.7. An experiment was conducted using this system. The results of this experiment are described in the following sections as follows: First, the results of the associated keyword generation are described in section 6. Then, the results of the document retrieval process are described in section 7.

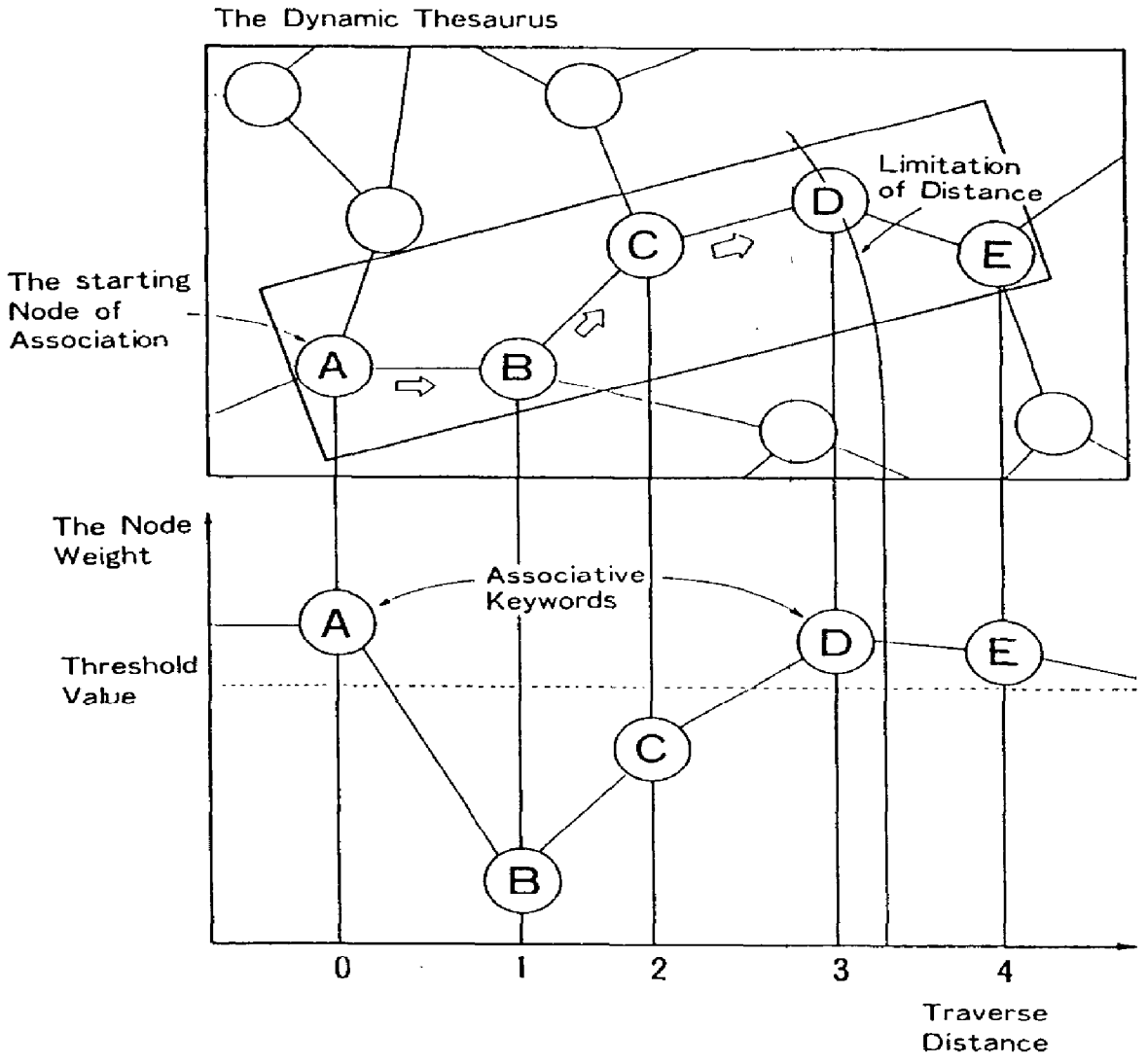

Fig. 6 Generation of Associated Keywords using links and node weight

6. Results of Associated Keyword Generation

Five newspaper articles were chosen as sample relevant documents for the experiment. These articles were then processed with the INDEXER system, which automatically extracts term information. A dynamic thesaurus was made from a static thesaurus and term information. Associated keywords generated from input keywords were evaluated against the keywords indexed in each relevant newspaper article. It was decided that if the associated keywords were a subset of the indexed keywords of a relevant document, then the associated keywords would be effective in

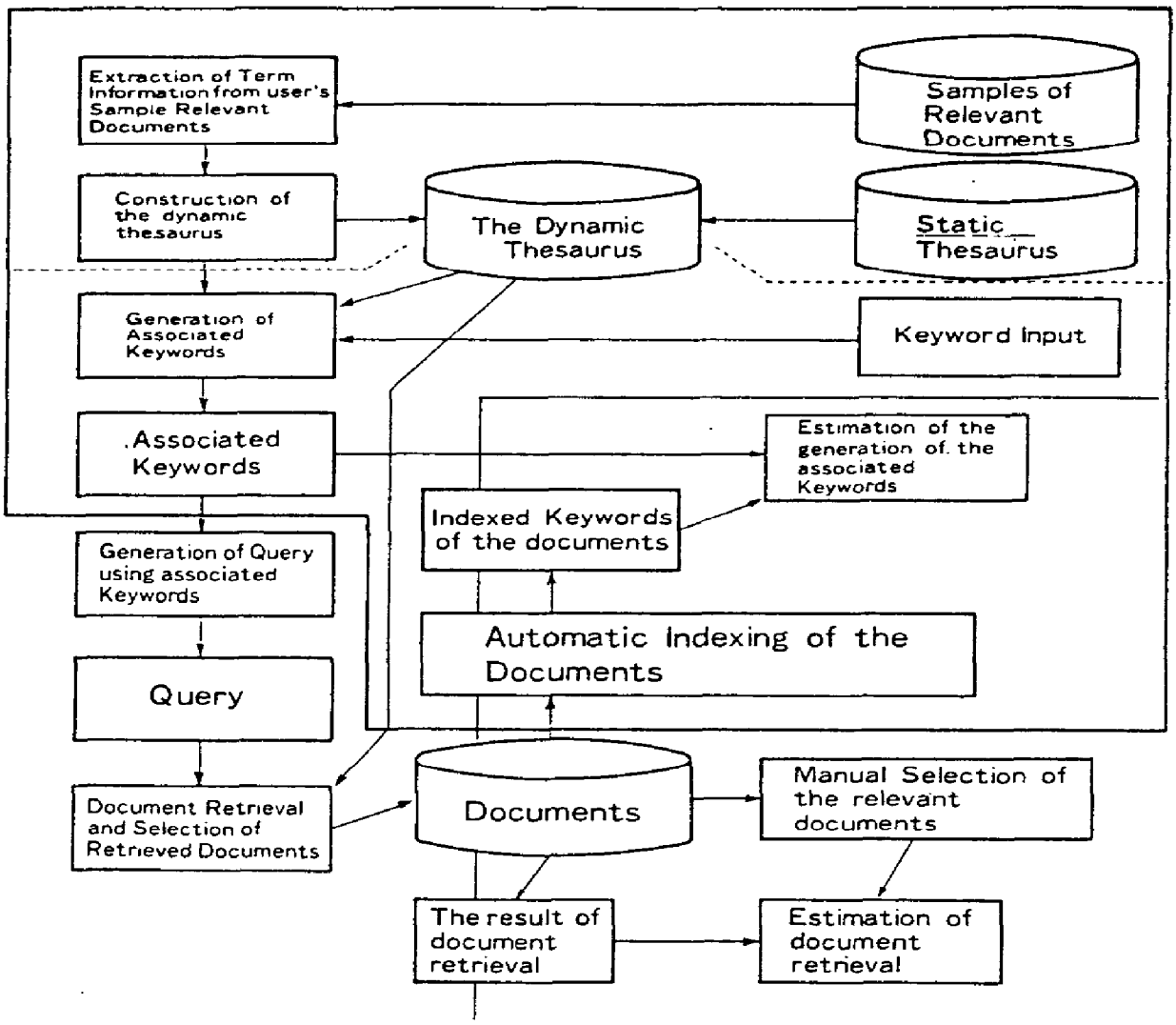

Fig. 7 Configuration of the AIRS Prototype

retrieving that document. The accuracy of associated keywords were measured by utilizing keyword recall rate and the keyword precision rate. Both a high keyword recall rate and a high keyword precision rate are necessary for accurate and effective document retrieval. The rates are defined as follows:

$$Kr = \frac{Nia}{Ni}, \text{ and}$$

$$Kp = \frac{Nia}{Na},$$

where Kr is the keyword recall rate; Kp is the keyword precision rate; Nia is number of the indexed keywords that are duplicated by the associated keywords; Ni is the number of indexed keywords; and Na is the number of associated keywords.

6.1 Experimental Environment

The prototype system processed the newspaper articles. Then, the thesaurus, which was made for retrieving these newspaper articles, was used as a static thesaurus. This static thesaurus contains about 8,000 terms.

6.2 Effect of Thesaurus Structure on Kp and Kr

Four kinds of thesauri were tested to find the most appropriate one for associated keyword generation. The thesauri tested are listed below.

Thesaurus A: The static thesaurus
(Nodes are unweighed; no co-occurrence links are used)
Thesaurus B: The dynamic thesaurus
(Nodes are weighed; no co-occurrence links are used)
Thesaurus C: The dynamic thesaurus
(Nodes are unweighed; co-occurrence links are used)
Thesaurus D: The dynamic thesaurus
(Nodes are weighed; co-occurrence links are used)

For all the dynamic thesauri the threshold value was set heuristically at 0.005. Table 1 and Fig.8 indicate the effect of each of these thesauri on Kp and Kr. At traversal distances greater than zero, Kr exceeds 70% if co-occurrence links are used. Kr is fairly constant and under 15% if co-occurrence links are not used or the traversal distance equals zero. Thus, links are very successful in accurately generating keywords. This is because the indexed keywords are extracted from those user relevant documents as one item in term information, and they are always linked with each other in the dynamic thesaurus. Thus, when one indexed keyword is inputted as a keyword, the other keywords which are indexed to the same document are always generated by traversing the links in the dynamic thesaurus.

Kp is about 30% to 50% when the thesaurus, whose nodes are weighed, such as B and D, are used, though Kp is 24% when only the input keywords are used. This means that the node weight is effective for improving Kp. Kp is highest for the type B thesaurus, where links are not used. This is because the type D thesaurus generates more and more keywords as the traversal distance increases. This suggests that the threshold value should increase with traversal distance.

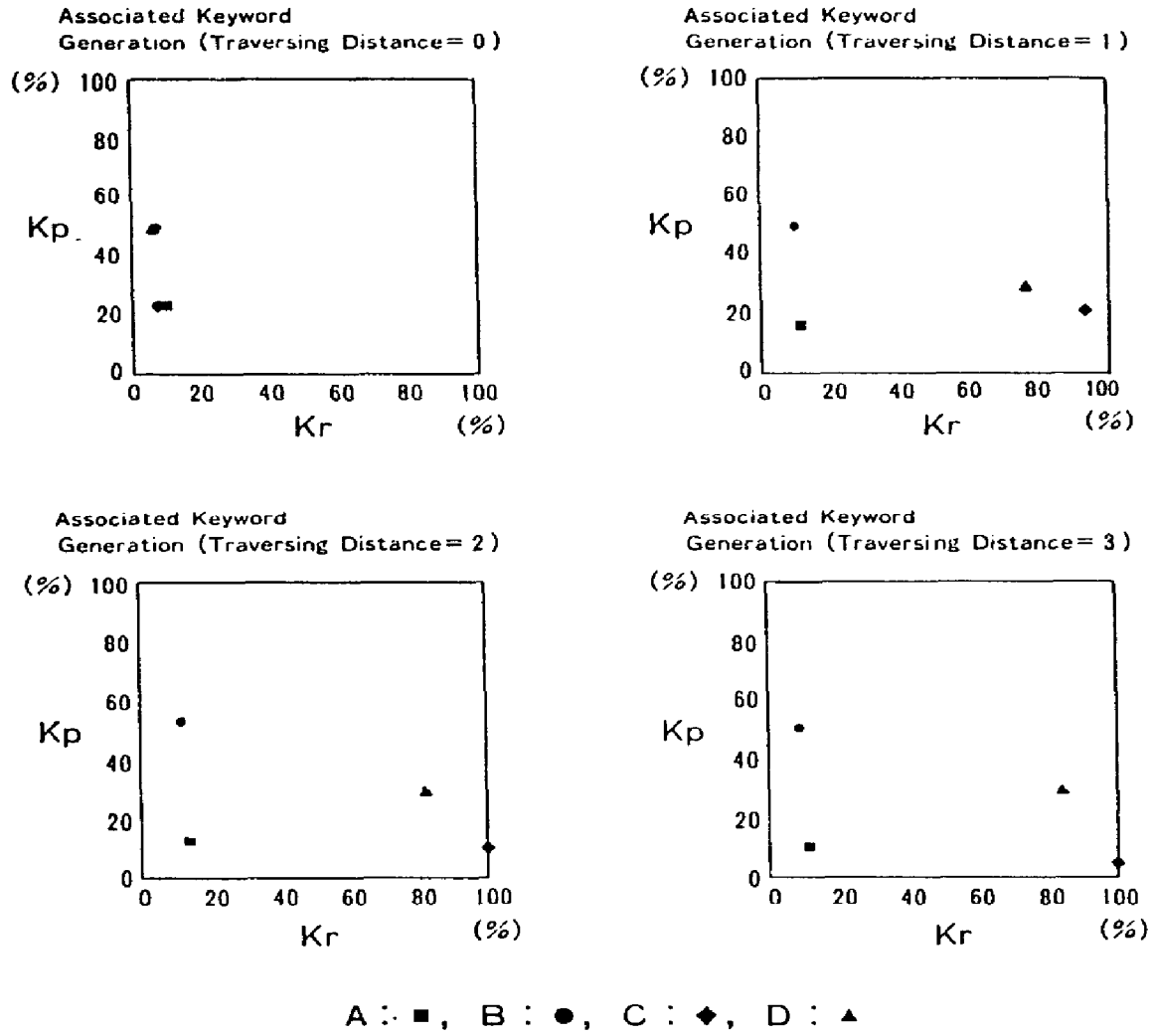

Fig. 8 Relationships between various kinds of Thesaurus Structures (A, B, C, D) and Associated Keyword Generation

. Table. 1 Effect of Thesaurus on Kr and Kp

Various Kinds of Thesaurus Structures	Usage of Generated Links	Usage of Node Weight	Traversing Distance 0		1		2		3	
			Kr	Kp	Kr	Kp	Kr	Kp	Kr	Kp
A : Static Thesaurus	×	×	9.60	24.00	12.82	15.24	14.36	12.68	12.95	10.10
B : Dynamic Thesaurus	×	○	7.18	51.67	10.39	50.33	11.93	53.33	9.91	50.00
C : Dynamic Thesaurus	○	×	9.60	24.00	94.00	21.78	100.00	9.28	100.00	5.81
D : Dynamic Thesaurus	○	○	7.18	51.67	77.08	29.18	83.08	28.85	83.38	28.85

6.3 Effect of Threshold Value on Kp and Kr

An experiment was conducted and the results are shown in Fig.9. In this experiment, a certain input keyword and the type D thesaurus were used. The results show that as the threshold value increases, Kr becomes smaller and Kp becomes larger. In AIRS, the same threshold value is applied regardless of the traversing distance. Both higher Kr and higher Kp could be achieved by introducing a variable threshold value that is dependent on the traversing distance.

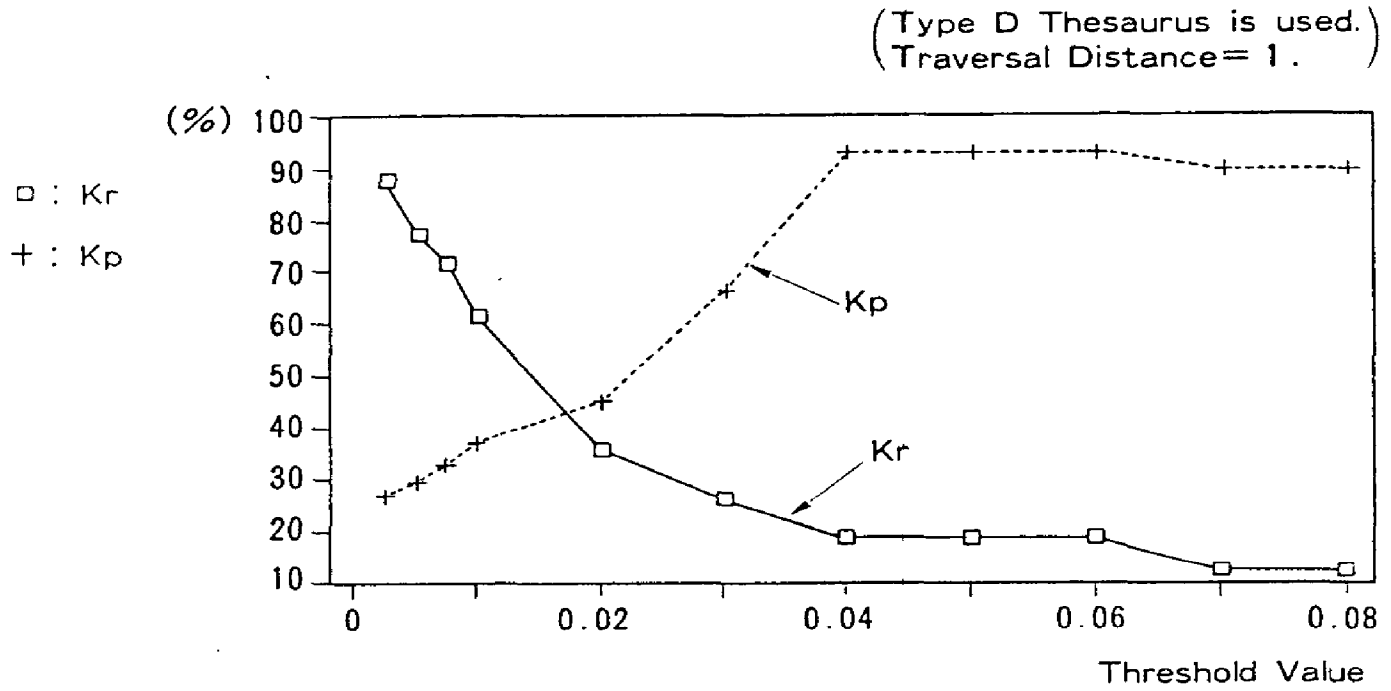

Fig. 9 Relationship between Threshold Value and Document Retrieval

7. Results of Document Retrieval Using Associated Keywords

A document retrieval experiment was performed using the associated keywords generated by AIRS. The results of this experiment are described in this section.

7.1 Experimental Environment

The experiment used a database of 163 newspaper articles. The experiment was conducted three times. The different documents were retrieved in each time. The numbers of relevant documents used to extract term information are; three for the first experiment; five for the second experiment; three for the third experiment . Keywords were inputted by a user, and then the associated keywords were generated by AIRS as described in section 3. The document retrieval was performed using these

associated keywords. The Boolean OR search strategy was adopted as a search strategy. Both a recall rate (Dr) and a precision rate (Dp) for the document retrieval were calculated. Three kinds of thesaurus structures were used in this experiment in order to find the most appropriate one among the three. They were the type A, C, and D thesauri. Two searchers attended the experiment.

7.2 Effect of Thesaurus Structure on Dr and Dp

An experiment was performed to find which type of thesaurus is most suited for document retrieval. The results of this experiment are shown in Fig.10 and Table 2. In Fig.10, the numbers attached to each point show which kind of thesaurus was used to get those particular data. The correspondence between the number and the type of thesaurus is as follows:

1: Thesaurus is not used.
2: Type A thesaurus is used.
3: Type C thesaurus is used.
4: Type D thesaurus is used.

The results show that when the type D thesaurus is used, both Dr and Dp are increased considerably. Thus, from the results of this experiment, the use of associated keywords generated by the type D thesaurus is the most effective for document retrieval.

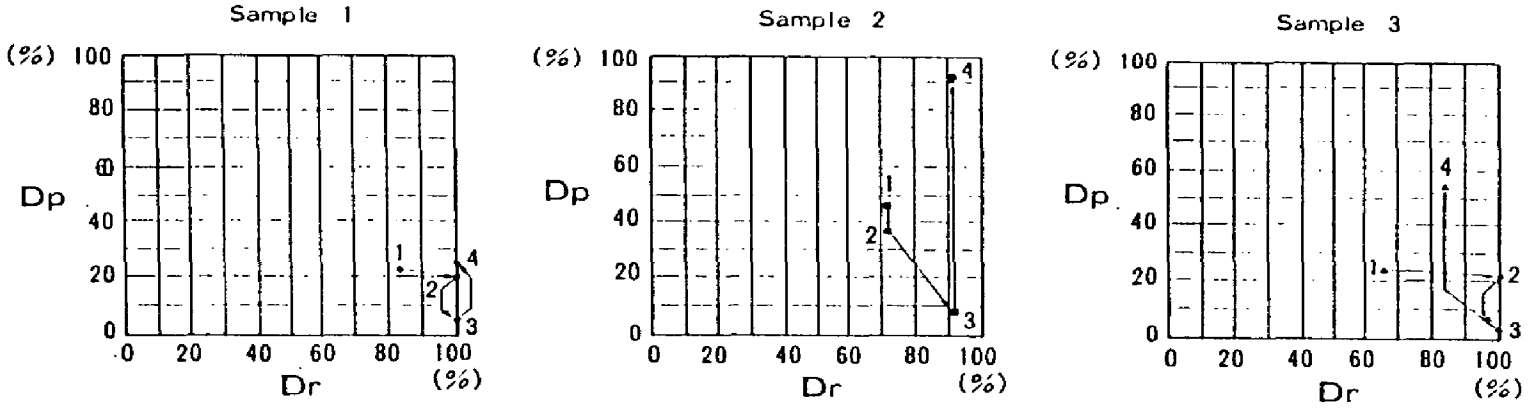

Fig. 10 The Relationship between various kinds of Thesaurus Structures and the corresponding results of document retrieval

Table. 2 Effect of Thesaurus on Dr and Dp

Various Kinds of Thesaurus Structures	Usage of Generated Links	Usage of Node Weght	Sample 1		Sample 2		Sample 3	
			Dr	Dp	Dr	Dp	Dr	Dp
1 : Retneval Using only User Input Keywords	—	—	83.33	20.83	72.73	47.06	66.67	28.00
2 : Static Thesaurus	×	×	100.00	20.00	72.73	38.10	100.00	24.00
3 : Dynamic Thesaurus	○	×	100.00	5.94	90.91	9.01	100.00	4.80
4 : Dynamic Thesaurus	○	○	100.00	24.00	90.91	90.91	83.56	55.56

8. Further Issues

The following are items that need to be researched in order to achieve higher Dr and Dp.

8.1 Construction of a Dynamic Thesaurus

8.1.1 Node weight calculating algorithm

The implemented system uses only the ranking of keywords, which are extracted from documents relevant to the user, as a means of measuring the importance of keywords. A more precise measurement would be possible if other information could be used such as keyword frequency, keyword location, syntactical information, and the time series information about a keyword.

8.1.2 Link generation and link weight calculating algorithm

The implemented system generates co-occurrence links whenever two keywords appear in the same relevant document. The generation of links should reflect the grammatical role of the keywords in the sentence, such as a subject-object relation. Furthermore, the link generation algorithm should generate various types of links, such as a cause-result link. Finally, all links should have a weight attached to them.

8.2 Associated Keyword Generation

During the traverse process in the dynamic thesaurus, the optimal node selection and optimal link selection should be calculated by using the node weight, the link weight, and so on.

References

[1] Hamill,K.A. and Zamora,A.: "The Use of Titles for Automatic Document Classification," Journal of the American Society for Information Science, Nov. 1980.

[2] Kimoto,H. Nagata,M. Kawai,A.:"Automatic Indexing System for Japanese Text," REVIEW of the Electrical Communications Laboratories, Vol.37, No.1, pp.51-56, 1989.

[3] Salton,G.: "Expert Systems and Information Retrieval," ASCM SIGIR Forum, Vol.21, No.3-4. 1987.

[4] Smeaton,A.F., Van Rijsbergen,C.J.: "Experiments on Incorporating Syntactic Processing of User Queries into a Document Retrieval Strategy," in Proceedings of the 11th ACM Conference on Research and Development in Information Retrieval, Presses Universitaires de Grenoble, pp31-51, 1988.