| | |
|---|---|
| Title | Automatic Indexing of Japanese Documents and its Application to Information Retrieval |
| Author(s) | 木本, 晴夫 |
| Citation | 大阪大学, 1993, 博士論文 |
| Version Type | VoR |
| URL | https://doi.org/10.11501/3072903 |
| rights | |
| Note | |

*The University of Osaka Institutional Knowledge Archive : OUKA*

https://ir.library.osaka-u.ac.jp/

The University of Osaka

# Automatic Indexing of Japanese Documents and its Application to Information Retrieval

Haruo KIMOTO

September 1993

Dissertation submitted to the Faculty of Engineering Science of Osaka University in partial fulfillment of the requirements for the degree of Doctor of Engineering.

Abstract

This thesis describes the automatic indexing of Japanese documents and its applications. The first part of this thesis describes the automatic indexing of Japanese documents and the second part describes its applications.

First, a new method for indexing individual Japanese documents is described, which effectively deletes extraneous words. These are words that have little relationship with the subject of a document. This method uses linguistic information, experts' indexing knowledge, and statistical information, such as word location and/or a word frequency. A system called INDEXER has been implemented using this method. This system has two main functions: deleting extraneous keywords among the keywords extracted by the free term method or the term control method, and ranking the keywords by evaluation scores. The evaluation criteria for automatic indexing methods are recall rate and precision rate. The conventional free term method has a recall rate of 70% and a precision rate of 10%. The new method achieves 50% for both rates. The 50% precision rate is a great improvement compared with the 10% precision rate of conventional methods. The new method provides an algorithm for ranking the extracted keywords. The experimental results show that 95% of the necessary keywords

are included in the top ten keywords ranked by this algorithm.

Many Natural Language Processing (NLP) techniques are used in this new method. The most important among them is a morphological analysis technique that achieves a very high preciseness in the analysis of Japanese sentences. This is done by using a large-scale semantic category dictionary as a pre-processor for automatic indexing.

There are several applications for this automatic indexing. Some of them are information retrieval, processing data in text databases, making teaching materials for teaching Japanese and etc.. The remainder of this thesis describes information retrieval.

The new information retrieval method described in this thesis is information retrieval with association, which incorporates a connectionist model in a dynamic thesaurus. This method is designed to discover and use the interests of a user so that the results of the document retrieval are more beneficial to that user. A prototype system was implemented, and called the Associated Information Retrieval System (AIRS). The basic concept of the new method is as follows. The system extracts a user's interests from the user's sample relevant documents as (a) keywords, (b) the degree of keyword importance in that document, and (c) the linkage between keywords in that document. The INDEXER system is used

to extract keywords and to determine the degree of keyword importance. Experiments were conducted to evaluate the new method and the results showed a high preciseness of document retrieval was achieved.

Chapter 1 is the introduction for this thesis. Chapter 2 is a preliminary for this thesis. Some technical terms and an introduction to NLP techniques are described. In Chapter 3, the necessity, the importance, the state of the art and the difficulties of automatic indexing of Japanese documents are described. Also, the purpose of this study is clarified in Chapter 3. Chapter 4 describes a new method for automatic indexing. This method selects and ranks keywords from each document. An analysis of the human way of indexing is also described. This analysis was made in order to obtain algorithms for the new method. A system was implemented using this new method and named the INDEXER system. Chapter 5 describes an evaluation of the new method through the evaluation of INDEXER. The services and the functions of INDEXER are also listed in Chapter 5. Chapter 6 presents applications of INDEXER. These include, using it in an information retrieval method with association, in the preparation of printed materials, in making Chinese character databases, in making electronic media such as CD-ROMs, and in indexing very large-scale text databases. Among these applications, is

an information retrieval method with association. This is quite a new method of information retrieval and one of the best applications of INDEXER. The application prototype system of an information retrieval method with association was implemented and named the AIRS system. This system is described in Chapter 7. The effect of the new method is described by evaluating AIRS. The conclusion and future work are described in Chapter 8.

List of Major Publications

Refereed Journal Papers

[1] Kimoto,H. and Iwadera,T., "Associated information retrieval system(AIRS) -Its performance and user experience," IEICE Trans. Inf. & Syst., Vol.E76-D, No.2, pp.274-283, Feb., 1993.

[2] Kimoto,H., "Automatic indexing and evaluation of keywords for Japanese newspapers," Transactions of IEICE Japan D-1, Vol.J74-D-1, No.8, pp.556-566, 1991(in Japanese).

[3] Iwadera,T. and Kimoto,H., "Automatic indexing of an integrated large scale text database and its evaluation," Journal of Japan Indexers Association, Vol.17, No.1, pp.15-24, Feb., 1993(in Japanese).

[4] Kimoto,H. and Ohba,T., "Automatic indexing technique," Journal of Japan Indexers Association, Vol.13, No.2, May, pp.1-12, 1989(in Japanese).

Refereed Papers in Conference Proceedings

[1] Sekine,J., Nakagawa,M., Kimoto,H. and Kurokawa,K., "A standard naming method of data elements using a semantic dictionary," Proceedings of the DEXA 92 (Database and Expert Systems Applications), Valencia, Spain, Springer-Verlag, Wien-New York, pp.167-172, Sep., 1992.

[2] Komori,S. and Kimoto,H., "Tagged corpus and its application to teaching Japanese," Proceedings of the Second International Conference on Foreign Language Education and Technology, Nagoya, Japan, pp.469-473, Aug., 1992.

[3] Iwadera,T. and Kimoto,H., "The effects of dynamic word network on information retrieval," Joint Conference of SPIE Conference on Applications of Artificial Neural Networks 3 and SPIE Conference on the Science of Artificial Neural Networks, Orlando, pp.362-369, Apr., 1992, (invited paper).

[4] Kimoto,H. and Iwadera,T., "Learning effect of a dynamic thesaurus in associated information retrieval," Proc. of the Second Workshop on Algorithmic Learning Theory (ALT91), Japanese Society of Artificial Intelligence (JSAI), Tokyo, October,

pp.47-57, 1991.

[5] Kimoto,H. and Iwadera,T., "A dynamic thesaurus and its application to associated information retrieval," Proc. of the IJCNN-91-SEATTLE, International Joint Conference on Neural Networks, Seattle, pp.I-19 - I-29, Jul., 1991.

[6] Kimoto,H., "Natural language processing and its application to Japanese database - the INDEXER system," Proceedings of the 3rd International Conference on Japanese Information in Science, Technology and Commerce, Vandoeuvre-les-Nancy, France, INIST, pp.447-460, May, 1991.

[7] Kimoto,H. and Iwadera,T., "Construction of a dynamic thesaurus and its use for associated information retrieval," Proceedings of 13th International Conference on Research and Development in Information Retrieval, Brussels, Belgium, Presses Universitaires de Bruxelles, pp.227-240, Sep., 1990.

[8] Kimoto,H., "An automatic indexing method using linguistic processing," Proc. of the First Annual Conference of JSAI, No.7-7, pp.389-392, 1987(in Japanese).

Acknowledgement

In completing this work, I have been fortunate to have received assistance from many individuals. I especially thank Professor Tohru Kikuno for his support, encouragement and guidance.

I am also very grateful to the members of my thesis review committee: Professor Tadao Kasami, Professor Nobuki Tokura and Professor Akihiro Hashimoto for their invaluable comments and helpful criticism of this thesis.

I thank my advisers in Message Processing Systems Laboratory of NTT Network Information Systems Laboratory, Mr. Minoru Nakamura, Dr. Yoichi Sakai and Dr. Nobuyoshi Terashima for giving me a chance to do this work.

I also thank Mr. Yasuo Sakama, Dr. Satoru Ikehara and Dr. Masaru Nakagawa of NTT Network Information Systems Laboratory for their support and encouragement throughout this work.

I thank Professor Tetsuya Ishikawa and Professor Yoshihumi Masunaga of the University of Library and Information Science for their discussion of the evaluation method of automatic indexing and information retrieval which ascertained the evaluation method adopted in this thesis.

I thank Mr. Mamoru Hiroki of Chunichi Newspaper Corporation for allowing me to use "News Thesaurus", which he edited, in the experiments described in this thesis and for discussions on how to make and apply thesauri.

I thank Ms. Akemi Haruyama for the discussion and the suggestions for associated keyword generation that were described in Chapters 6 and 7.

I thank Professor Vijay V. Raghavan and Professor Edward Fox for their comments about the dynamic thesaurus and the description of the experiments of associated information retrieval in Chapters 6, 7 and 8.

Finally, I especially thank Mr. Toshiaki Iwadera for his discussion, suggestions and assistance in completing this work. He made AIRS and did many experiments with me. I also thank Mr. Yoshinori Kishida and Ms. Naoko Takahashi for helping in the experiments and the evaluations of AIRS.

A list of captions

Fig. 1 Segmentation of a compound noun.
Fig. 2 Giving readings along Chinese character.
Fig. 3 Flow of research and development of automatic indexing.
Fig. 4 Difference between keywords and non-keyword distribution in texts.
Fig. 5 General operating flowchart of INDEXER.
Fig. 6 Keyword selection results.
Fig. 7 Effect of deletion algorithm.
Fig. 8 Relationships between query and relevant documents.
Fig. 9 Conventional model and AIRS model.
Fig. 10 General process flow diagram of AIRS.
Fig. 11 Construction of dynamic thesaurus using term information.
Fig. 12 Co-occurrence of keywords in documents.
Fig. 13 Generation of links.
Fig. 14 Node weight calculation.
Fig. 15 Generation of associated keywords using links and node weight.
Fig. 16 Example of giving readings along Chinese characters.
Fig. 17 Schematic of associated information retrieval.
Fig. 18 Distribution of nodes in dynamic thesaurus.

Fig. 19 Sample documents and corresponding results of document retrieval.
Fig. 20 Threshold value and results of document retrieval.
Fig. 21 Result of document retrieval.

Table 1 Semantic connection rules in compound nouns.
Table 2 Comparison of automatic indexing systems.
Table 3 Result of analysis for selecting keywords.
Table 4 Service menu of INDEXER.
Table 5 Comparison of INDEXER with other systems.
Table 6 Evaluation of INDEXER ranking function.
Table 7 Efficiency of AIRS.

Contents

Chapter 1. Introduction.....1
1.1 Introduction.....1
1.2 Outline of the thesis.....8
Chapter 2. Preliminaries.....11
2.1 Technical terms.....11
2.2 Natural language processing (NLP) technique.....12
2.2.1 Necessity.....12
2.2.2 State of the art.....12
2.2.3 High precision morphological analysis.....13
Chapter 3. Automatic indexing of Japanese documents...19
3.1 Necessity.....19
3.2 Importance.....20
3.3 State of the art.....20
3.4 Difficulties.....26
3.5 Purpose of this study.....27
Chapter 4. A new method for automatic indexing.....29
4.1 Overview.....29
4.2 An analysis for a new method.....29
4.3 Extraction of keyword candidates.....33
4.4 Keyword selection without rank.....35
4.4.1 Linguistic information.....35
4.4.2 Experts' knowledge.....38
4.4.3 Statistic information.....40
4.5 Keyword selection with rank.....43
4.6 Originality of the new method.....44

Chapter 5. Evaluation of the new method.....45
5.1 The INDEXER system.....45
5.1.1 Services and functions of INDEXER.....45
5.1.2 Procedures of INDEXER.....47
5.2 Experimental data.....50
5.3 Evaluation.....50
5.3.1 Keyword selection without rank.....50
5.3.2 Keyword selection with rank.....53
Chapter 6. Applications of INDEXER to information retrieval with association.....56
6.1 Information retrieval.....56
6.2 Problems in information retrieval.....56
6.3 A new method for information retrieval.....59
6.3.1 Key idea.....59
6.3.2 Basic concept of the new method.....59
6.3.3 Algorithm.....68
6.4 Originality of the new method.....75
6.5 Other applications of INDEXER.....76
6.5.1 Making paper materials.....77
6.5.2 Making electronic media.....80
6.5.3 Making Chinese character databases.....80
6.5.4 Indexing a very large-scale text database...81
Chapter 7. Evaluation of AIRS.....83
7.1 The AIRS system.....83
7.2 Experimental data.....85
7.3 Evaluation.....85

7.3.1 Evaluation criteria.....85
7.3.2 Construction of dynamic thesaurus.....86
7.3.3 Document retrieval using AIRS.....89
7.4 Other expected effects of AIRS.....97
Chapter 8. Conclusions.....99
8.1 Summary.....99
8.1.1 Automatic indexing.....99
8.1.2 Information retrieval with association.....100
8.2 Future works.....100
8.2.1 Automatic indexing.....100
8.2.2 Information retrieval with association.....101

References.....103

# Chapter 1. Introduction

## 1.1 Introduction

This thesis describes the automatic indexing of Japanese documents and its applications. The first part of this thesis describes the automatic indexing of documents and In the second part describes its applications. There are two ways of retrieving information from databases of documents or papers. One way is to use bibliographical indexes, such as the author's name, title, publisher's name, and etc.. Another way is to use subject indexes, such as classification numbers and keywords. Both of these indexes are attached to each document in a database for retrieval. Classification numbers are used to locate documents on the shelves of libraries, and keywords are used to show the contents of the documents. An automatic indexing method for bibliographical indexes and subject indexes is needed to retrieve documents in large-scale databases. For automatic indexing of bibliographical indexes, an automatic recognition method of pages indicating where that data are written and a recognition method of characters are needed. For automatic subject indexing, a content analysis method of documents is necessary.

In the real world, subject indexing is done by human indexers. They read, analyze and index documents using terms in thesauri. Those indexes are called manual

indexes. They are supplemented with those words extracted automatically from titles and abstracts of the documents, which are called free words. However, the volume of documents to be indexed is growing very rapidly, and human indexing will not be able to meet the growing needs of indexing. Also, free words can not be used as subject indexes because they are not indexed to documents by analyzing the contents of the documents.

Automatic indexing by computer has been studied for many years. There are two basic automatic indexing models. One model collectively extracts keywords from a set of texts, such as newspaper articles, technical papers, and so on. The other model indexes one article at a time.

To extract keywords from a set of articles, a statistical approach is generally adopted, such as an analysis of the frequency of word occurrence[Luhn 1957][Salton 1975][Nagao et al. 1976]. But this approach is still in the experimental stage for indexing Japanese documents.

For indexing individual articles, the free term method is used[Kinukawa et al. 1982] [HAPPINESS]. The free term method extracts all the nouns in the text except those contained in a user-defined dictionary, which is called a stop word dictionary. As a result, all of the necessary keywords in the text are extracted but many extraneous words are also extracted. Extraneous

words are words that have little relation with the content of the document (see Chapter 2). The percentage of extraneous words extracted is generally around 90%. Considerable memory is wasted on these extraneous keywords, and in retrieving documents, many meaningless articles are retrieved.

One way to solve the problem of meaningless articles being retrieved is to use a thesaurus. This way is called the term control method[Nakazono et al. 1984] [Abe 1985]. In this method, words are first extracted using the free term method. Then, from these, the words that match the thesaurus are selected as keywords. However, even the term control method still extracts many extraneous words.

The free term method and the term controlled method have already been commercialized. These methods are usually used in two ways. One way is to use only the free term method and the other way is to use both methods at the same time.

Some commercial automatic indexing systems, which give the subject index, have already been developed for documents written in Japanese [HAPPINESS] [Hayakawa et al. 1992]. But none of them does a precise analysis of Japanese sentences. Japanese sentences are very difficult to analyze using computers because the words are not separated, while in English the words are separated. To solve this problem, a high precision morphological

analysis technique, which is one of the NLP techniques, is necessary.

Many NLP techniques are used in INDEXER. NLP is an interesting research field and there are many expected applications such as machine translation (MT), fact extraction from text databases, man-machine dialogue and others. NLP is expected to play a vital role in the recognition and synthesis of voices or characters. Many NLP research projects are being carried out all over the world. Some of the NLP applications are a sentence reading system [Miyazaki et al. 1986], a sentence checking support system [Ohara et al. 1991] and an automatic indexing system [Kimoto 1987d][Kimoto 1991a][Kimoto 1991b].

Sections 2.2.2 and 2.2.3 describe the state of the art of the NLP technique and a morphological analysis technique [Miyazaki 1984], respectively. This morphological analysis technique achieves a very high preciseness in the analysis of Japanese sentences. This is accomplished by using a large-scale semantic category dictionary as a pre-processor for automatic indexing.

The purpose of this study is to develop a method that provides the subject index to the documents automatically. The target language is Japanese, and the documents are newspaper articles, but not limited to only these types of documents.

This paper proposes a new method for indexing indi-

vidual articles, which effectively deletes extraneous words. This method uses linguistic information, experts' indexing knowledge, and statistical information, such as word location analysis and/or word frequency analysis.

A system called INDEXER has been implemented using this method [Kimoto 1987a][Kimoto 1987b][Kimoto 1987c][Toriyama 1992]. INDEXER has two main functions: deleting extraneous keywords from the keywords extracted by the free term method or the term control method, and ranking the keywords by evaluation scores.

Experimental results show that INDEXER deletes more than 80% of the extraneous keywords. The evaluation criteria for automatic indexing systems are the recall rate and the precision rate. The recall rate is defined as the number of automatically indexed keywords that match the manually indexed keywords, divided by the number of manually indexed keywords. The precision rate is the number of automatically indexed keywords that match the manually indexed keywords, divided by the number of automatically indexed keywords. The free term method has a recall rate of 70% and a precision rate of 10%. The INDEXER system achieves 50% for both rates. The 50% precision rate is a great improvement. This means that one out of every two words among the keywords extracted by INDEXER is significant to the indexed text. INDEXER has a keyword ranking function, which ranks the extracted keywords. The user can then use the ranked

keywords as desired, selecting the top five keywords or the top ten keywords, and so on. Experimental results show that 95% of the necessary keywords are included in the top ten keywords as ranked by this function.

There are several applications for automatic indexing. Some of them are information retrieval, making teaching materials for teaching Japanese and etc.. The rest of this thesis describes an application for information retrieval.

In the field of information retrieval, the development of word-processors, optical disk filing systems, and computer networks enable us to use large-scale databases. However, before we can advance database systems, there are two problems that must be solved: document storing and document retrieval.

For document storing, an automatic document classification system and an automatic indexing system have already been developed [Hamill et al. 1980][Kimoto et al. 1989][Kimoto 1991a][Kimoto 1991b].

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 [Salton 1987] and natural language interface systems [Smeaton et al. 1988][Akiyama 1988][Kanou et al. 1991].

There are many difficulties with the above mentioned methods. Development of an expert system requires a rule base for each application. On the other hand,

there are many difficulties involved with natural language interface techniques. The first of these is that it is difficult to make precise syntactic and semantic analysis of a user queries automatically [Smeaton et al. 1988]. Only the morphological analysis technique can be used from a practical point of view. The second is the cost of domain knowledge acquisition. Domain knowledge includes such items as a domain model, a world model and a large word dictionary for each domain [Akiyama 1988]. The third problem is that much dialogue between a user and a system is required for making a practical user model [Kanou et al. 1991].

This thesis proposes a new method that incorporates a 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 method was implemented into a system called the Associated Information Retrieval System (AIRS).

The basic concept of this new method is as follows. AIRS extracts a user's interest from the user's sample documents as (a) keywords, (b) the degree of keyword importance in that document, and (c) the linkage between keywords in that document. INDEXER [Kimoto et al. 1989][Kimoto 1991a][Kimoto 1991b] is used for extracting keywords and the degree of keyword importance.

The whole mechanism of AIRS is as follows. Key-

words, keyword importance (ranking) and the co-occurrence relation of keywords in a document constitute what we refer to as term information. The dynamic thesaurus is made from the static thesaurus using this term information. The associated keywords are generated from the input keyword of a user and the dynamic thesaurus. These associated keywords are used to retrieve documents that precisely fit the user's interest. The evaluation results of the associated keyword generation, and the relationships between the data structure of the dynamic thesaurus and the result of the document retrieval [Kimoto et al. 1990] are described herein.

The expected effects of the new method are as follows:(1) Associated keywords, which reflect a 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 a user's prior interest is possible.

Experiments were conducted to evaluate the new method and the experimental results showed a high preciseness of document retrieval using AIRS.

## 1.2 Outline of the thesis

Chapter 2 discusses some preliminaries for this

thesis. Some technical terms and an introduction to the Natural Language Processing (NLP) techniques are described. In Chapter 3, the necessity, the importance, the state of the art and the difficulties of automatic indexing of Japanese documents are described, and the purpose of this study is clarified. Chapter 4 describes a new automatic indexing method. This method selects and ranks keywords from each document. An analysis of the human way of indexing is also described. This analysis was conducted in order to obtain algorithms for the new method. A system was implemented using this new method and named the INDEXER system. Chapter 5 describes the effect of the new method by evaluating INDEXER. The services and the functions of INDEXER are listed in this chapter. Chapter 6 presents applications of INDEXER. These include using it in an information retrieval method with association, in the preparation of printed materials, in making Chinese character databases, in making electronic media such as CD-ROMs, and in indexing very large-scale text databases. Among these applications, the information retrieval method with association is quite a new method of information retrieval and it is one of the best applications of INDEXER. The application prototype system of the information retrieval method with association was implemented and named AIRS. This system is described in Chapter 7. The effect of the new method is described by evaluating of AIRS. The

conclusions and future works are described in Chapter 8.

# Chapter 2. Preliminaries

This chapter describes and explains some technical terms and a technique, which are necessary for automatic indexing and information retrieval. The technical terms are explained in Section 2.1 and the Natural Language Processing(NLP) techniques are explained in Section 2.2.

## 2.1 Technical terms

(i) Manual keyword

Manual keywords are keywords indexed by a human expert indexer. Indexing by human is called manual indexing.

(ii) Extraneous keyword

Among the keywords extracted using an automatic indexing system, the ones that do not match the manual keywords are called extraneous keywords.

(iii) Thesaurus

In this thesis, a thesaurus is defined as a collection of terms that are used for information retrieval. When these terms are used for indexing documents or used in queries of information retrieval, they are called keywords. As a logical structure, thesaurus has a broader term and narrower term relation and a synonym relation between terms. The number of keywords in a thesaurus depends on each application.

Chapter 2

## 2.2 Natural language processing (NLP) techniques

### 2.2.1 Necessity

It is necessary for automatic indexing systems to extract nouns from documents because keywords are almost always nouns. Sometimes verbs or adjectives are required as keywords. They are used in cases as when one wants to retrieve documents using verb keywords such as "move," "fly," etc., or when using adjective keywords such as "red," "large," etc.. Segmenting words in a sentence, giving readings to them and distinguishing their parts of speech constitutes the morphological analysis technique. There are other techniques in NLP, such as, syntactic analysis and so on. The syntactic analysis technique is useful in selecting important words from documents. So, these NLP techniques are necessary for precise automatic indexing. The next section describes NLP as a whole.

### 2.2.2 State of the art

A great amount of NLP research is now being done. This research includes work in the areas of automatic translation from Japanese into English and vice versa, automatic sentence checking and automatic indexing. NLP is comprised of morphological analysis, syntactic analysis, semantic analysis and discourse analysis. In the state of the art of NLP, sentence checking systems and automatic indexing systems, with a morphological analy-

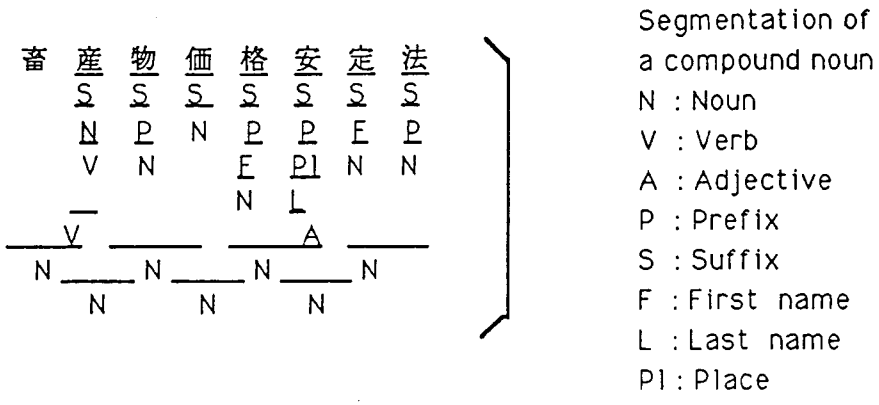
sis technique, are being used in the real-world. Other techniques such as semantic analysis or discourse analysis, are not yet actually being used in real-world applications[Ishikawa 1992][Nagao 1992][Yokoi 1993].

### 2.2.3 High precision morphological analysis

This section describes a high precision morphological analysis technique that is now being used in some NLP systems such as an automatic sentence checking system and etc..

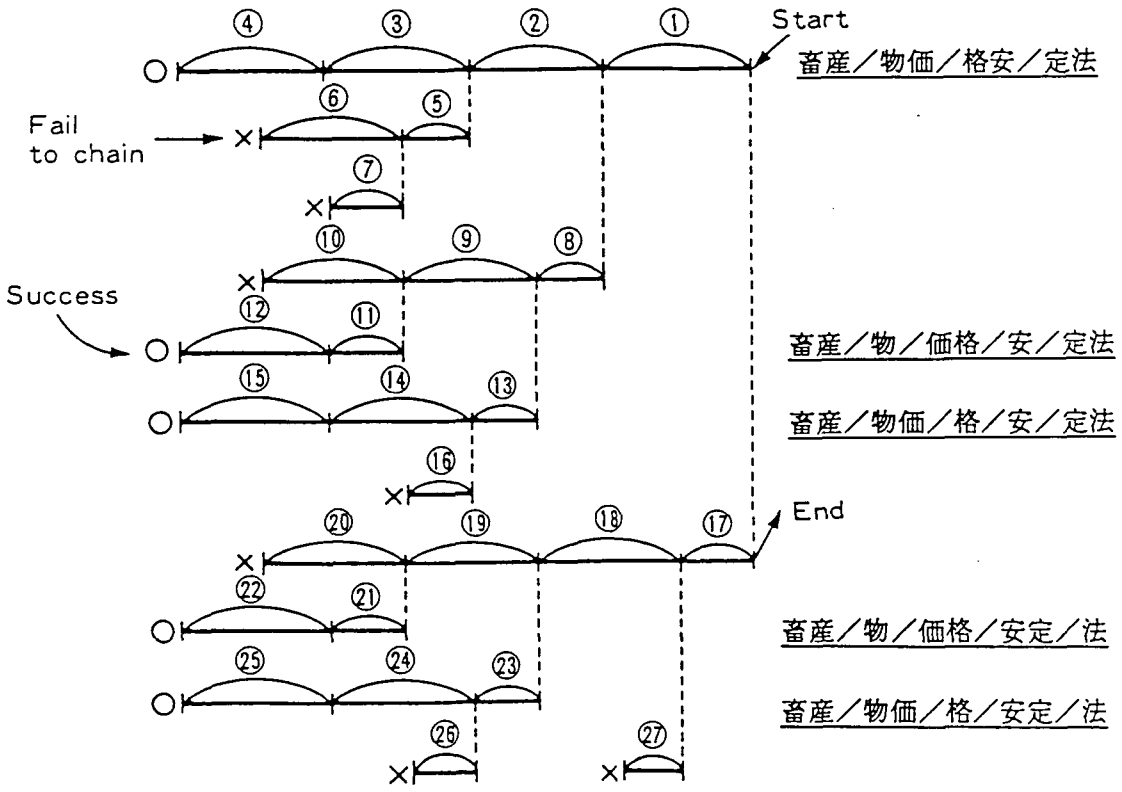
There are no spaces between words in Japanese sentences. In order to achieve a high degree of precision in the analysis of Japanese sentences, it is therefore necessary to make a correct sentence segmentation between words and to correctly designate each word as a verb, noun, etc. (This is called morphological analysis.) For this purpose, a large scale dictionary database was developed, which contains approximately 430,000 words and around 60 columns for each word. This dictionary contains grammatical and semantic connection rules between words, in addition to entry words, parts of speech and readings contained in dictionaries that are used in ordinary machine translation systems. These connection rules make it possible to achieve a high degree of precision in word segmentation. An example of correct segmentation of a compound noun, using these connection rules, is shown in Fig. 1(a),(b). In the

figure, five candidates are chosen as correctly segmented using grammatical connection rules, and these candidates are checked using semantic connection rules. The morphological analysis technique establishes segmentations that are grammatically and semantically correct using grammatical connection rules and semantic connection rules.




(a) Possible segmentations.

Fig. 1 Segmentation of a compound noun.




(b) Segmentations using grammatical connection rules.

Fig. 1 Segmentation of a compound noun.

Chapter 2

Some semantic connection rules are shown in Table 1. The total number of rules is 15. Each rule was written in the form of "a kind of a part of speech" and "-" and "a kind of a part of speech." This form means that the first kind of a part of speech connects with the

Table 1 Semantic connection rules in compound nouns.

| No. | Semantic connection rule | | Examples |
|---|---|---|---|
| 1 | Prefix-Number | | 約 10、 第 八 回 |
| 2 | Number-Suffix | | 二 本、 50 パーセント |
| 3 | Number-Suffix | | 50kg 強、 数 % 台 |
| 4 | Number-Suffix | | 100 未満、 10 以下 |
| 5 | Pronoun -Suffix | Place | 東京 駅、 関東 平野 |
| | | Name | 平野 副 社長、 一郎 君 |
| | | Organization | 三井 信託 銀行 |
| | | Other pronouns | 明治 時代、 アイヌ 人 |
| 6 | Prefix-Post | | 美濃部 前 都 知事 |
| 7 | Last name-First name | | 加藤 一二三 |

second one in this order. For example, No.1 rule in Table 1 means that prefix connects with number in this order. Examples of readings provided (in the Japanese phonetic syllabary) along with Chinese characters (pictographs of Chinese origin) using these semantic connection rules are shown in Fig. 2.

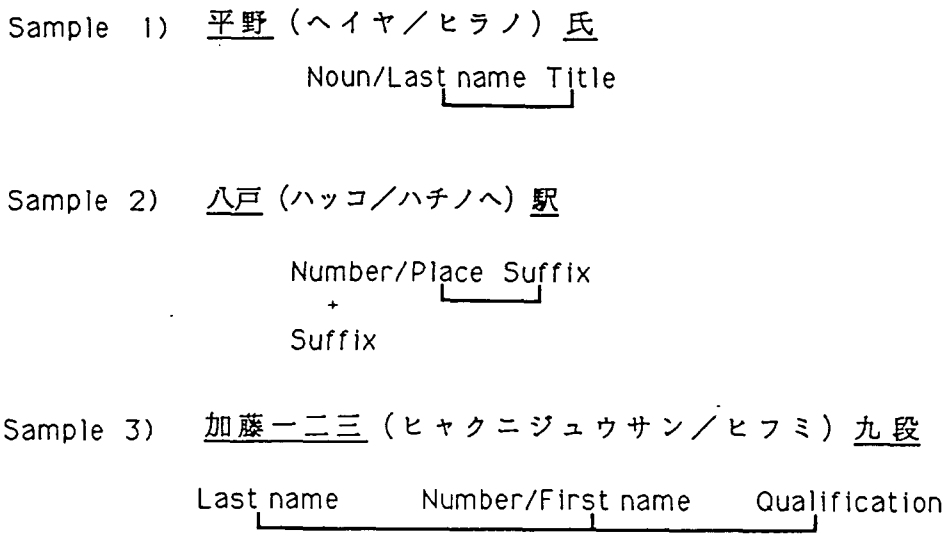



Fig. 2 Giving readings along Chinese character.

In Sample 1 in Fig. 2, the reading of 平野 can be either ヘイヤ(heiya) or ヒラノ(Hirano). For ヘイヤ(heiya), the semantic attribute of 平野 is 'general word', and for ヒラノ(Hirano), the semantic attribute of 平野 is 'last name'. As the word following 平野 is 氏, whose semantic attribute is 'title', the semantic connection rule of 'last name-title' is applied and ヒラノ(Hirano) is given to 平野 as its correct reading.

Automatic indexing systems use outputs of the morphological analysis system, such as the part of speech and readings of each segmented word.

# Chapter 3. Automatic indexing of Japanese documents

## 3.1 Necessity

Automatic indexing is very necessary in retrieving text databases, such as newspaper articles, legal articles, patent papers and technical papers. These databases are growing very rapidly in commercial fields and in in-house corporate database services fields, and the precise retrieval of documents in these databases is very necessary. Nowadays, many people are engaged in the indexing task and so it is very expensive to index. In some cases, there are about 70 peoples doing indexing in one database service company. So, it would be impossible to enumerate the total number of peoples who are engaged in indexing in the database services field, including commercial database services and in-house database services.

As more than one person is engaged in indexing, the indexing result is heterogeneous and this causes the inaccurate retrieval of documents. The time requirement for indexing is very short because people want to retrieve documents as soon as they can. As a result, in many cases, indexing is done at midnight or in the early morning just after new documents, such as newspapers, are published.

An automatic indexing technique resolves these problems. It eliminates the number of peoples engaged in

indexing, and it indexes homogeneously because a computer program always does the indexing the same way for all documents and it runs regardless of the time.

## 3.2 Importance

A high precision automatic indexing technique is applicable to other applications. In the state of the art, described in Chapter 1 and in Section 3.4, the automatic indexing technique in the real world selects keywords regardless of the meanings of keywords in the document. It would be very significant if keywords were selected according to the meanings of words in documents. If meaningful words were selected as keywords, the preciseness of automatic indexing would be very high. Furthermore, the selection of meaningful words would prove to be very successful in supporting many techniques such as automatic document classification and routing, automatic document summarizing, automatic understanding of texts and automatic extraction of knowledge from texts.

## 3.3 State of the art

This section describes the history of the research and development of automatic indexing. The research of automatic indexing started with the development of the computer. The first system was made by H.P. Luhn in the late 1950s when computers were called electronic data

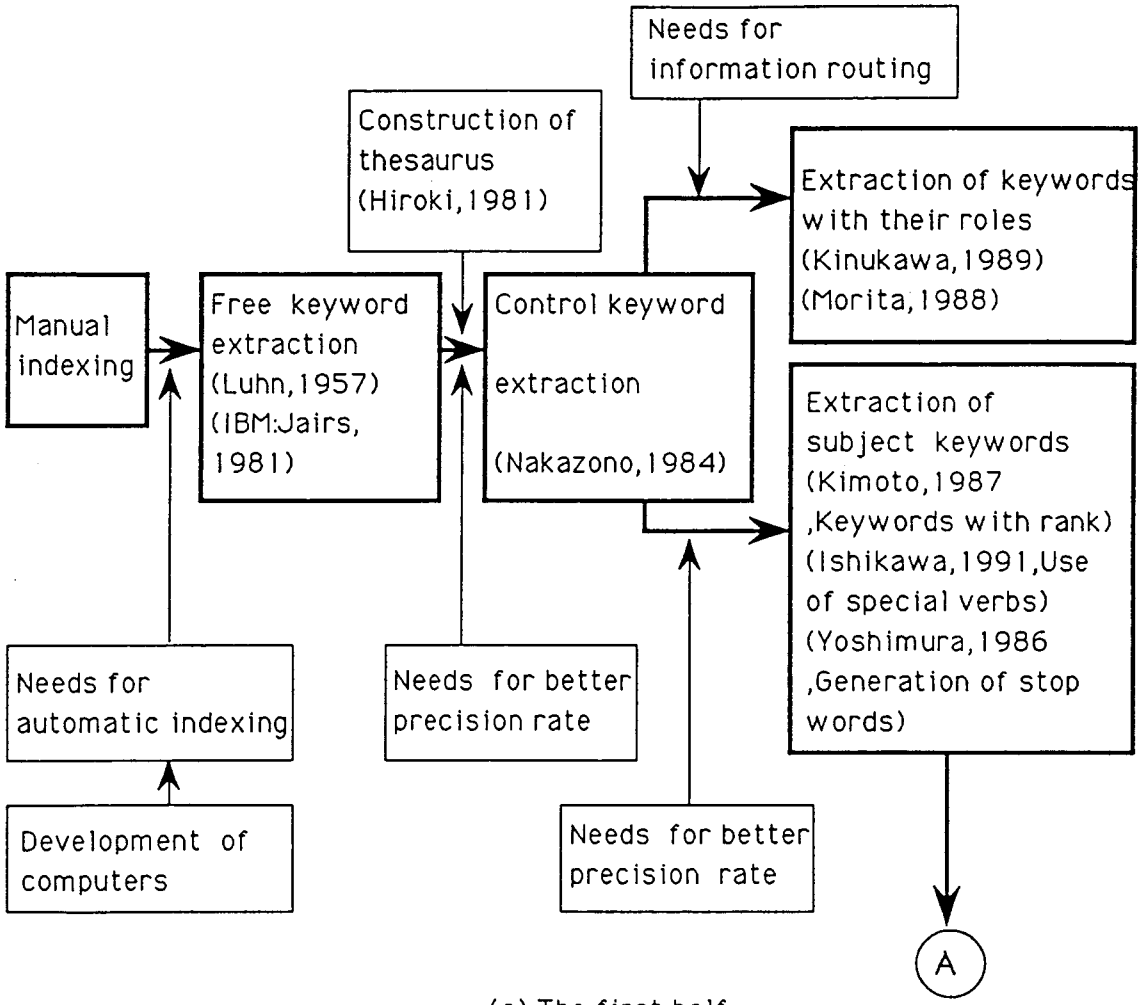
processing systems (EDPS)[Luhn 1957]. At that time, the processing speed was not as high as it is now, and the memory and storage capacities were very low. In the following paragraphs, automatic indexing methods are described in the historical order of the research and development of the technique. A flow of the research and development of automatic indexing is shown in Fig. 3(a),(b).

(i) Automatic indexing using a frequency analysis of words

H.P. Luhn developed a system that would extract words and count the frequency of these words in documents and when a frequency of a word is larger than a threshold value, the word is selected as a keyword [Luhn 1957]. Some words, such as "of," "a" and "the" appear frequently in documents, but they are never keywords. So, they are called stop words and deleted from keywords. The disadvantage of this system was that it could not extract compound nouns as keywords, because, for example, "of" is deleted from compound nouns. S.Futamura made a further study of this method using a statistical analysis technique of the frequency of words [Futamura et al. 1987].

(ii) Automatic indexing using a keyword list

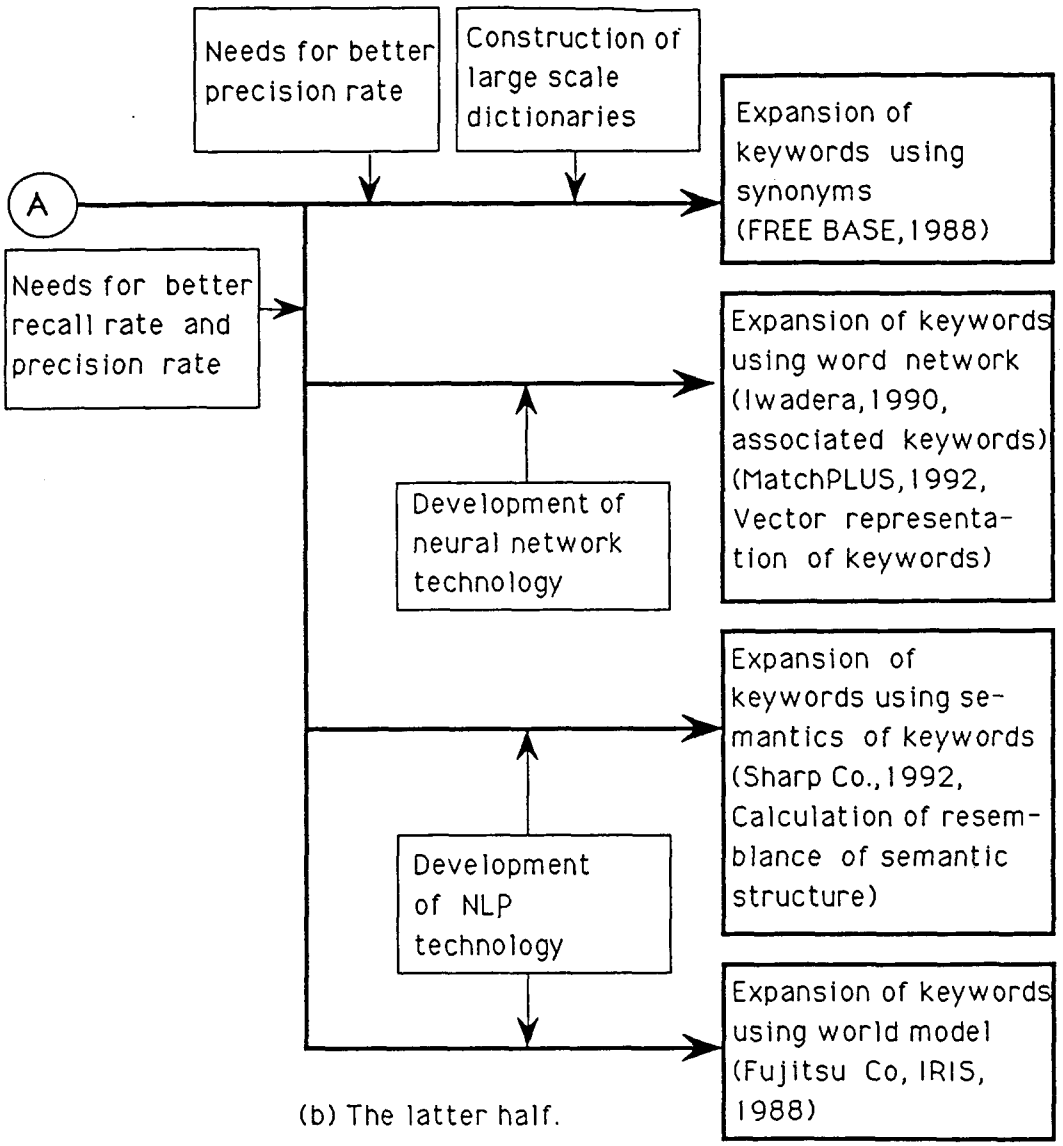
In this method, a large list of keywords, including




(a) The first half.

Fig. 3 Flow of research and development of automatic indexing.

(with the change of needs, development of hardware, development of related technologies and their years)




(b) The latter half.

Fig. 3 Flow of research and development of automatic indexing.

(with the change of needs, development of hardware, development of related technologies and their years)

compound noun keywords and single word keywords, are prepared and used. When a word in a document matches one of the keywords on the list, the word is selected as a keyword. This method requires a large storage capacity, but is not as expensive as storing a large list of keywords. This amount of storage became available with the development of the storage techniques and the method became commercialized. The systems that use this method for indexing Japanese documents are, HAPPINESS by the Heiwa Jyoho Center [HAPPINESS] and Free Base by the MC Word Center. The commercialized systems are divided into the free term method and the term control method as described in Chapter 1. The term control method uses a thesaurus to control terms, but the free term method does not. These two methods are the two basic models of automatic indexing these days. These systems must maintain a large list of keywords in order to guarantee the completeness of the list against the change of keywords or newly appearing keywords over time.

(iii) Automatic indexing by sentence analysis

This method uses the NLP technique, especially the syntactic analysis technique, for analyzing structures of sentences in a document in order to extract structured keywords. Structured keywords mean that each keyword has a role in a sentence or in a set of sentences and a set of keywords with a role constitutes struc-

tured keywords. Examples of roles are, an agent, an objective, a place, a method and so on. H.Kinukawa developed a system that extracted a set of structured keywords, which is equivalent to the so called five W's and one H, from Japanese newspaper articles [Kinukawa et al. 1982]. Y.Morita also reported extracting structured keywords [Morita,Y. 1988]. The output of these systems are structured keywords, therefore, these systems are not commonly used because usually people use keywords that represent the whole content of a document. These keywords are so called subject keywords or subject index.

There has been a lot of research done on extracting subject keywords. K.Yoshimura reported a prototype system of automatic extraction of technical terms from Japanese scientific documents [Yoshimura et al. 1986]. This system uses the elements of a stop word dictionary and connection rules of these elements in order to generate strings of these elements. These strings of stop word elements are deleted from the keyword candidates that were extracted using a morphological analysis of sentences in a document. Shibata proposed using particular kinds of verbs as keys for extracting subject indexes [Shibata et al. 1987]. These researches have a common problem, in that words other than subject keywords are extracted together with subject indexes.

Research has been conducted on a machine-aided sub-

ject indexing system [Ishikawa 1991]. The input of this system is an abstract of a research paper in the information processing field, and the output is phrases and sentences that contain subject keywords. This system is based on a method that uses some specific verbs that express the author's intention of writing the paper, as keys for extracting the phrases and sentences mentioned above. Although the evaluation results of this system are good, it is not sufficient for automatic indexing.

## 3.4 Difficulties

One of the biggest problems of automatic indexing is that it is very difficult for a computer program to understand the meanings of words in documents. There are many papers on natural language analysis such as syntax analysis, semantic analysis, natural language understanding and so on. But almost all of them are at the research level. Few are really effective in the real world, and they are used in small limited fields, such as, weather forecast translation from English to French and vice versa. We need to analyze the meanings of words in documents.

The commercial methods used nowadays are, the free term method and the term control method. Both methods extract many extraneous words (see Section 2.1), and one approach to get a better indexing method is to distinguish necessary keywords from extraneous words. For this

purpose, a sentence analysis technique and a document analysis technique, which uses a Natural Language Processing (NLP) technique, would be effective. In Chapter 4, a new method for automatic indexing, which uses an NLP technique, is described.

## 3.5 Purpose of this study

The purpose of this study is to find an automatic indexing method whose indexing preciseness is nearly equivalent to that of human indexing. The preciseness of human indexing is shown in Table 2 together with those of some existing automatic indexing methods. One problem that must be solved is finding out how human indexers index documents. To accomplish this, an analysis of human indexing was conducted using documents and keywords indexed to those documents by human indexers, and from this analysis a new method was developed. The new method uses an NLP technique, some rules from indexing experts' knowledge, as well as a statistical analysis technique for automatic indexing.

Table 2 Comparison of automatic indexing systems.

| Criteria / System | Recall rate | Precision rate |
|---|---|---|
| Free term | 70% | 10% |
| Term control | 65% | 20% |
| * Manual indexing | 70-80% | 70-80% |

* Correlation of indexing results for different expert indexing

# Chapter 4. A new method for automatic indexing

## 4.1 Overview

This Chapter proposes a new method for automatic indexing that effectively deletes extraneous keywords. The proposed method uses linguistic information, experts' indexing knowledge, and statistical information, such as, word location information and/or word frequency information. In order to find a method, an analysis of extraneous keywords among keyword candidates was conducted. A simulation was made using a method that was obtained from the result of the analysis, and the simulation experiment showed good automatic indexing results. Therefore INDEXER was implemented using the method [Kimoto 1987a][Kimoto 1987b][Kimoto 1987c]. This system has two main functions: deleting extraneous keywords among the keywords extracted using the free term method or the term control method, and ranking the keywords by evaluation scores.

## 4.2 An analysis for a new method

In order to develop a precise automatic indexing system, it is necessary to select keywords among keyword candidates. In other words, it is necessary to delete extraneous keywords. An analysis was made in order to find algorithms for selecting necessary keywords and algorithms for deleting extraneous keywords.

Chapter 4

For this analysis, control keywords were extracted as keyword candidates from the samples of 30 newspaper articles using the control term method. A thesaurus for indexing newspaper articles [Hiroki 1981] was used. This thesaurus has about 9,000 words and logical relations among these words include the broader term relation, the narrower term relation and the synonym relation. Among the extracted 219 keyword candidates, 58 were necessary keywords and the other 161 were extraneous keywords.

The following algorithms were obtained through the analysis. 1. Deletion of parallel words, 2. Deletion of modifier words, 3. Deletion of broad-category words, 4. Deletion of words that express time or place, 5. Generation of grouping words from the keyword candidates of the same category, 6. Deletion of words that express titles of people. These algorithms are explained in the following paragraphs.

(i) Deletion of parallel words

This algorithm deletes keyword candidates that are expressed in parallel in a sentence. When keyword candidates of A, B and C are extracted from a sentence such as; "I gave my diamonds to A, B, C and others.", these keyword candidates are deleted and they do not become keywords.

(ii) Deletion of modifier words

This algorithm deletes modifier words from the keyword candidates. When "dream" is a keyword candidate extracted from a sentence; "My dream of super express became true.", "dream" is deleted from the keyword candidates because it is a modifier word of "super express".

(iii) Deletion of broad-category words

This algorithm deletes broad-category words from the keyword candidates. This is because broad-category words have too wide a range of meanings to be keywords for retrieving some specific documents.

(iv) Deletion of words that express time or place

This algorithm deletes words that express time or place from the keyword candidates. The time, such as hour or minute, are seldom used as keywords. Also, "Apple" will not be a keyword in its original meaning of fruit when it is extracted as a keyword candidate from a sentence such as; "I came from Apple Town."

(v) Generation of grouping words from the keyword candidates of the same category

This algorithm generates grouping words as keywords from the keyword candidates of the same category. For example, "political party" will be generated as a key-

word when there are keyword candidates of "Liberal Democratic Party," "Socialist Party" and "Communist Party."

(vi) Deletion of words that express titles of people

This algorithm deletes words that express titles of people from keyword candidates. For example, in the sentence "The director of Nippon Telegraph and Telephone corporation, Mr. Sato visited the USA.", "Nippon Telegraph and Telephone corporation," "Sato" and "USA" are important words, and "director" is not so important. In this sentence, "director" is used as a title of a special person, and it is not used for its proper meaning. This sentence is seldom retrieved using the keyword "director." So, in such a case, "director" is deleted from keyword candidates.

The simulation results using these algorithms on the same 30 newspaper articles are shown in Table 3. These results show that the algorithms of "Deletion of parallel words," "Deletion of modifier words" and "Deletion of broad-category words" are effective in deleting the extraneous keywords from keyword candidates. This is to say, it is possible to delete extraneous keywords by identifying these features, such as parallel words, for each keyword candidate in a document. The algorithms mentioned above are described in

detail in Section 4.4.1.(i)-(iii).

It was also ascertained in another analysis that "Emphasized words in the sentence" and "A word pair of broad-category words and a narrower word in a thesaurus" are effective in selecting keywords. These algorithms were made through the analysis of the results of human indexing. This analysis is not described in Table 3. These algorithms are described in detail in Section 4.4.1(iii), (iv).

The algorithms using experts' knowledge and statistic information are also useful for selecting keywords and are described in sections 4.4.2 and 4.4.3, respectively.

## 4.3 Extraction of keyword candidates

Before selecting keywords, it is necessary to extract keyword candidates from a document. Keyword candidates are nouns excluding stop words, prefixes and suffixes. For extracting nouns, the high precision morphological analysis technique, described in Section 2.2.3, is used. After extracting nouns, stop words, which are in the user defined dictionary, are excluded from these nouns. Then, prefixes and suffixes are also excluded. The rest of the nouns are called free keywords. Among free keywords, the words that match the words in a thesaurus are control keywords. From either free keywords or control keywords, keywords are select-

Table 3 Result of analysis for selecting keywords.

| No. | Algorithms for deleting extraneous keywords and for selecting necessary keywords | The number of deleted extraneous keywords in the simulation on 30 newspaper articles | The number of deleted necessary keywords in the simulation on 30 newspaper articles |
|---|---|---|---|
| 1 | Deletion of parallel words | 16 | 0 |
| 2 | Deletion of modifier words | 28 | 1 |
| 3 | Deletion of broad-category words | 16 | 1 |
| 4 | Deletion of words which express time or place | 3 | 1 |
| 5 | Generation of grouping words from the keyword candidates of the same category | 0 | 0 |
| 6 | Deletion of words which express titles of people | 1 | 0 |

ed and ranked. The processes described in this section correspond to steps 1 through 4 in the procedures of INDEXER described in section 5.1.2.

## 4.4 Keyword selection without rank

In this section a keyword selection algorithm is described that selects necessary keywords from the free keywords or the control keywords that were extracted by the pre-processor in INDEXER. The algorithms described in Section 4.4.1 correspond to step 5 in Section 4.4.2, to step 6 in Section 4.4.3 and to step 7 in the procedures of INDEXER in section 5.1.2, respectively.

### 4.4.1 Linguistic information

(i) Deletion of parallel words

Parallel words are words that appear in the following form in a text:

(a) A and B
(b) A or B
(c) A, B

Parallel words are less emphasized than single words in texts. About 200 newspaper articles were analyzed to clarify whether or not parallel words are generally significant to the text. The results of the analysis showed that almost no parallel words qualify as

keywords. Parallel words are thus deleted from the keyword list.

(ii) Deletion of modifier words

Modifier words are words that qualify other words. For example,

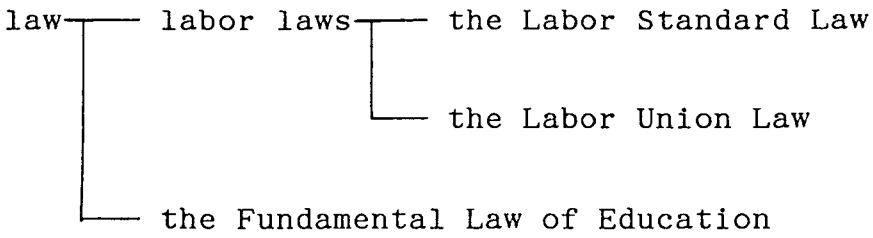
- "the world" is a modifier word
  in "the music of the world"

In general, modifier words are less significant than other words, and thus do not qualify as keywords. There are a number of cases, however, where modifier words are still significant. For example, where the modifier word implies a person and so on. Also, modifier words are important by definition if they are included in the user-defined significant word dictionary.

(iii) Deletion of broad-category words

Broad-category words encompass other narrower, more specific words in the thesaurus. Examples of broad-category words are, "law," "commerce," "information," etc.. More specific words that might be subsumed under the broad-category word "law", for example, would be "bill," "act," and so on. Thus, broad-category words represent more general concepts than the narrower words. The concepts expressed by the broader words are so wide that too many newspaper articles would have to be retrieved if broad-category words were permitted as key-

words. Therefore, broad-category words are generally not fit to serve as keywords. Narrower words, on the other hand, represent specific concepts, so they do qualify as keywords. In INDEXER, the News Thesaurus [Hiroki 1981] published by Kinokuniya has been used for broad-category word identification. The following is an example of the relations between broad-category words and narrower words.




Broad-category words do however, become keywords in special cases, such as the following.

(a) A broad-category word and a narrower word form a significant word pair in a text

There are many cases where a broad-category word and a narrower word appear linked together in a text. These word pairs often have a greater significance to the text than either of the words standing alone. This significance becomes clear from the fact that the broad-category word includes the meaning of the narrower word and the narrower word specifies the notion of the

broader word. This pair of words can thus become keywords as a pair. Word pair examples might include:

"airport" —— "Tokyo Airport"
"fish" —— "Tuna"

(b) Proper Nouns as Broad-category Words

In the thesaurus used in INDEXER, proper nouns can also be categorized as either broad or narrow. For example, "Japan" is a broad-category word encompassing "Tokyo." Although "Japan" is a broad-category word, it is significant because "Japan" is a proper noun. Proper nouns become keywords even if they are broad-category words. Proper nouns are obviously very important for newspaper article retrieving.

(iv) Emphasized words as keywords

It is apparent that some words are emphasized or stressed by neighbor words or phrases. For example, in the phrase, "this desk of mine" the word "desk" is being emphasized by its association with the word "this." Emphasized words that are preceded by such functional words as "this," "that," and "at first" are thus selected as keywords.

4.4.2 Experts' knowledge

This section describes the algorithm that generates

keywords using expert indexers' knowledge.

(i) Patterns of indexing

There are fixed indexing patterns that expert indexers use. Several samples of indexing patterns are listed below.

(a) There are regular keywords for special articles, such as an article named "COLUMN." The column name and the author's name are established as keywords.

(b) Conference names are divided into two parts. That is, "conference" and the subject name of the conference. For example, "conference" and "artificial intelligence application."

(ii) Strong association of keywords

Some keywords are strongly associated. For example, the phrase "equal opportunity employment regardless of sex" is strongly associated with "equal rights for both sexes." There are several kinds of associations. One word to one word, one word to several words, and so on. These associations are used in two ways. One way is to generate keywords that are not in the text by association with a keyword that has already been extracted from the text. The other way is to select words as keywords by association with already selected keywords when those words might otherwise be deleted from the keyword list due to parallel, modifier, or broad-category word dele-

tion.

For example, the word "company" has an association with the words "settlement of accounts" and "personnel affairs." Thus, if "company" is selected as a keyword, "settlement of accounts" is also fixed as a keyword by association with "company," although "settlement of accounts," as a broad-category word, might be deleted. "Personnel affairs" is also fixed as a keyword, it may be a parallel word.

(iii) Generating classification words as keywords

Company names often appear in newspaper articles, especially in the economic section. Companies are classified as a particular group of business or industry. So when a company name appears, the corresponding industry name is also generated as a keyword. The generation of the industry name is achieved using a classification table. For example,

"Ford" generates "Motor Industry"

4.4.3 Statistic information

(i) Deletion by word location

There is a tendency for more important words to appear early in texts. Newspaper articles can be divided into the headline and the body. Almost all significant words in the newspaper article appear in the headline

and in the first and the second sentences of the body. Significant words are seldom first introduced in the rest of the article.

An analysis was conducted to determine the boundary between the significant part and the less significant part of newspaper articles. The sample consisted of 200 newspaper articles that had already been manually indexed. The results of the analysis are shown in Fig. 4. Almost all manually indexed keywords appear in the first half of the article, while non-keywords are randomly distributed throughout the article. From the analysis of the distribution of manually indexed keywords in the newspaper articles, it can be determined that the significant portion of the newspaper article is from the top to the 86th character. Adopting this result, words that appear after the 86th character are deleted from the keyword list.

(ii) Choosing keywords by word frequency

It is also well established that the more frequently a word appears, the more important it is. The importance of the word in the text can be measured, by how frequently it appears. A certain critical frequency divides the important words from the unimportant. This critical frequency depends on the length of the text. The length of ordinary newspaper articles ranges from 200 to 1000 characters. Heuristically, we decided to

select words that appear more than 4 times in a text as keywords. In the INDEXER system, this critical frequency can be set by the user.

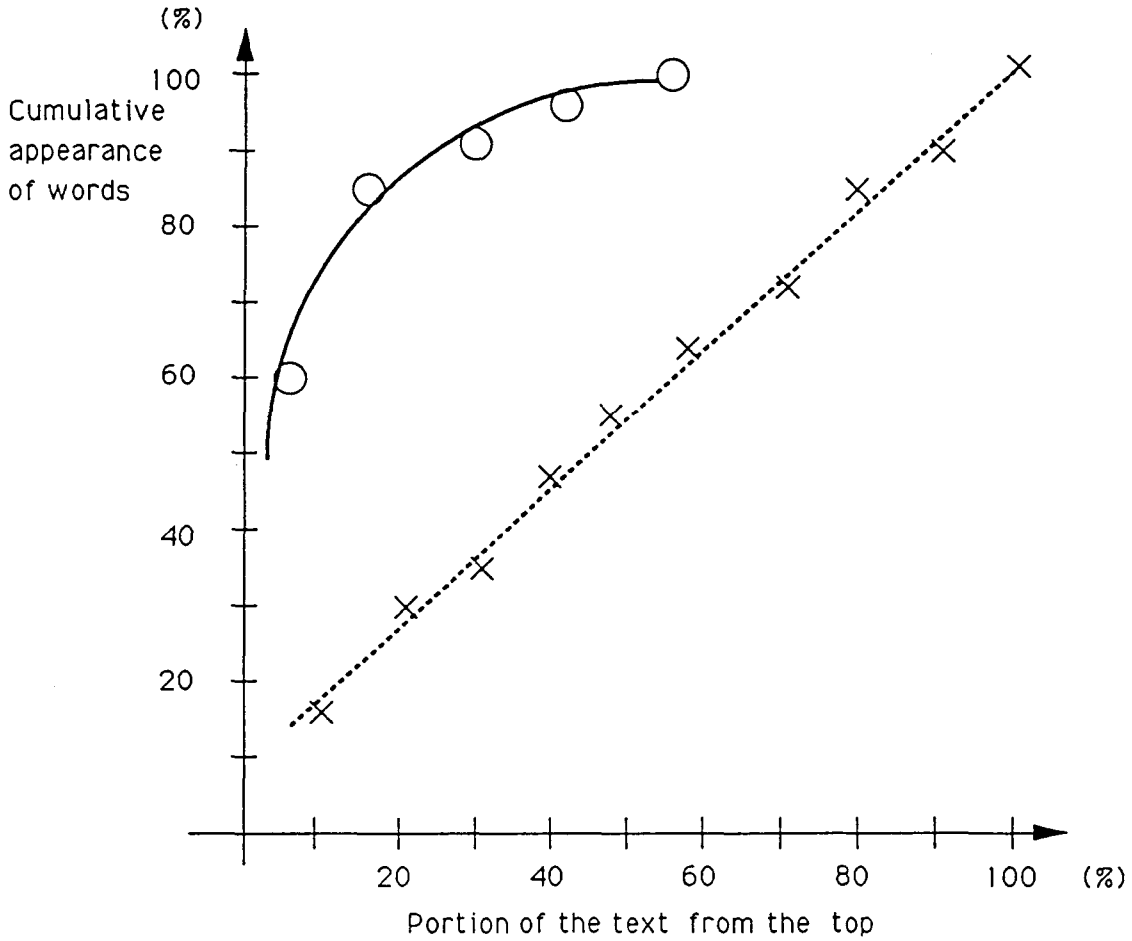



Fig. 4 Difference between keywords and non-keyword distribution in texts.

(iii) Combined use of word location and word frequency

Combining word location and word frequency is also effective in determining keyword status. An algorithm that uses such a combination was introduced. This algorithm evaluates words in the headline in the following way. A word that appears in the first part of a headline becomes a keyword if it appears more than twice in the body. Similarly, a word that appears after the second part of the headline becomes a keyword if it appears more than three times in the body.

## 4.5 Keyword selection with rank

The algorithms described above were used for keyword selection. The results of the experiment show that while most of the extraneous keywords were effectively deleted, a number of necessary keywords were also deleted at the same time. The results of the selection of keywords by INDEXER show that about 4 keywords were extracted for each newspaper article, including 2 necessary keywords on the average. The number of deleted necessary keywords was 0.65 per article.

For users who require greater precision with no necessary keywords being deleted, INDEXER ranks keywords according to their importance in a text. The importance of the keyword in a text is measured using the same algorithms that are used to select keywords. For this measurement, each algorithm is assigned a certain

weight. For example, the parallel word deletion algorithm is given a weight of -30. When a keyword fits a particular algorithm, that keyword gets points equal to the algorithm's weight. Each keyword is evaluated by each algorithm and given a total cumulative score. Keywords are then ranked according to these total scores. The user can select keywords from this ranked keyword list. Thus, necessary keywords are never deleted or lost. The algorithm described in this section corresponds to step 8 in the procedures of INDEXER described in Section 5.1.2.

## 4.6 Originality of the new method

A new method for automatic indexing was described in the above sections. The original idea was that this method would use linguistic analysis, experts' indexing knowledge, and statistic information such as word location analysis and/or word frequency analysis for indexing individual articles. Especially, the parallel words deletion algorithm, the modifier words deletion algorithm and other linguistic analyses are quite original. The new method effectively deletes extraneous words using these algorithms with the statistical analyses of words, which is a well known classical way of analysis.

# Chapter 5. Evaluation of the new method

## 5.1 The INDEXER system

### 5.1.1 Services and functions of INDEXER

To evaluate the new method, a system named INDEXER was developed as an application of NLP to Japanese text database processing. It is an automatic indexing system for Japanese text databases. INDEXER utilizes the morphological analysis technique and the technique of providing readings as described in Section 2.2.3.

The service menu of INDEXER is listed in Table 4. One major function of INDEXER is automatic extraction of keywords from documents for database retrieval. In addition, INDEXER automatically provides readings along with Chinese characters. INDEXER provides a variety of keyword extraction levels. These are as follow;

Level 1: Nouns (verbs, adjectives and segmented compound nouns are optional) are extracted as keywords.

Level 2: Free keywords are extracted, i.e. only stop words are deleted from noun keywords.

Level 3: Control keywords are extracted.

Level 4: Frequency and location information of keywords are extracted.

Level 5: Keywords are selected using frequency and/or location of keywords in the documents and

other information.

Level 6: Keywords are ranked using frequency and/or location and other information.

Table 4 Service menu of INDEXER.

| No. | Service menu |
|---|---|
| 1 | Segmentation of sentence into words and giving the part of speech of words |
| 2 | Noun extraction |
| 3 | Free keyword extraction using a stop word dictionary |
| 4 | Control keyword extraction using a thesaurus |
| 5 | Selection of keywords by linguistic analysis and statistical analysis |
| 6 | Evaluation of keywords by linguistic analysis and statistical analysis |
| 7 | Giving 'Kana' for Kanji |
| 8 | Segmentation of a compound noun into words |
| 9 | Extraction of verbs and adjectives |
| 10 | Getting statistic information about keywords, such as frequency |

The INDEXER user can use any level of keyword extraction. The merits of using INDEXER are as follow;

1: It is possible to sharply reduce the manpower required for indexing.
2: It is possible to shorten the time required for indexing.
3: It is possible to homogenize the indexing result using a computer system. (It is difficult to index homogeneously when several peoples are indexing jointly.)
4: It is possible to use INDEXER to make book indexes.

5.1.2 Procedures of INDEXER

The general operating principles of INDEXER are described here as applied to newspaper article indexing. The system operates as follows.

Step 1: Nouns are extracted from the input newspaper article.
Step 2: The words in the user-defined dictionary, so called stop words, are deleted.
Step 3: Prefixes and suffixes, such as "ex" in "ex-president", are removed from these nouns.
Step 4: Next, the nouns that match the thesaurus or the significant word dictionary are extracted as

keywords and compiled as a keyword list.

Step 5: Keywords are then linguistically analyzed. Comparatively less important words, such as A, B and C in the parallel expression "A, B, C" are deleted from the keyword list. Modifier words are also deleted and broad-category words, which have a wide range of meaning, are also deleted. Emphasized words are selected as keywords. Four algorithms, described in Section 4.4.1, achieve these respective functions.

Step 6: A human indexer experts' knowledge is used to either select or generate keywords. Three algorithms, described in Section 4.4.2, achieve these functions.

Step 7: Keywords are selected by analyzing keyword location and/or keyword frequency in the text. Three algorithms, described in Section 4.4.3, achieve these functions.

Step 8: The keywords are assigned a certain weight by each of the algorithms in Step 5 through Step 7 and are ranked according to the sum of these weights.

A general flowchart of the system is presented in Fig. 5.

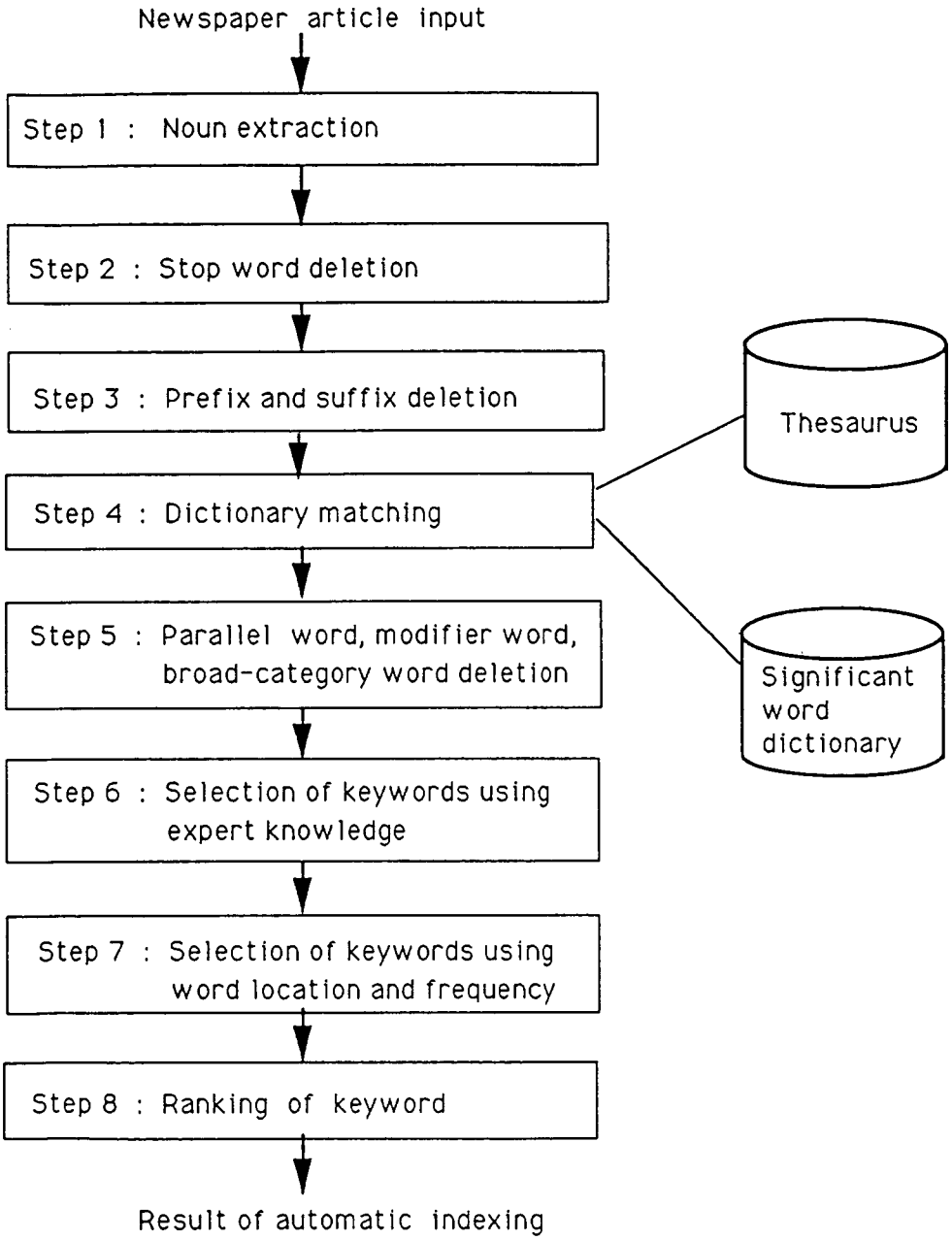



Fig. 5 General operating flowchart of INDEXER.

Chapter 5

## 5.2 Experimental data

INDEXER was implemented on a mini-computer and evaluated. Sample newspaper articles used for the evaluation were chosen from all the main fields of a daily newspaper: politics, economics, industry, sports, home affairs, domestic affairs, and so on. The number of articles sampled was 200. The sample articles were manually indexed, and the manually indexed keywords formed the basis for our evaluation of the system.

## 5.3 Evaluation

### 5.3.1 Keyword selection without rank

The evaluation criteria for automatic indexing systems are the recall rate, the precision rate, and the deletion ratio of extraneous keywords. The recall rate and the precision rate were defined in Section 1.1 (Introduction). The free term method has a recall rate of 70% and a precision rate of 10%. INDEXER achieves 50% for both rates. The 50% precision rate is a great improvement. It means that one out of two words among the keywords extracted by INDEXER is significant to the indexed text. Table 5 shows the recall rate and the precision rate achieved by INDEXER compared with those achieved by other systems in existence and those of manual indexing.

The deletion ratio of extraneous keywords is defined as the number of extraneous keywords deleted using

INDEXER, divided by the total number of extraneous keywords before keyword selection. Figure 6 shows the results of the keyword selection. The deletion ratio was

Table 5 Comparison of INDEXER with other systems.

| Criteria / System | Recall rate | Precision rate |
|---|---|---|
| Free term | 70% | 10% |
| Term control | 65% | 20% |
| INDEXER | 50% | 50% |
| * Manual indexing | 70-80% | 70-80% |

* Correlation of indexing results for different expert indexing

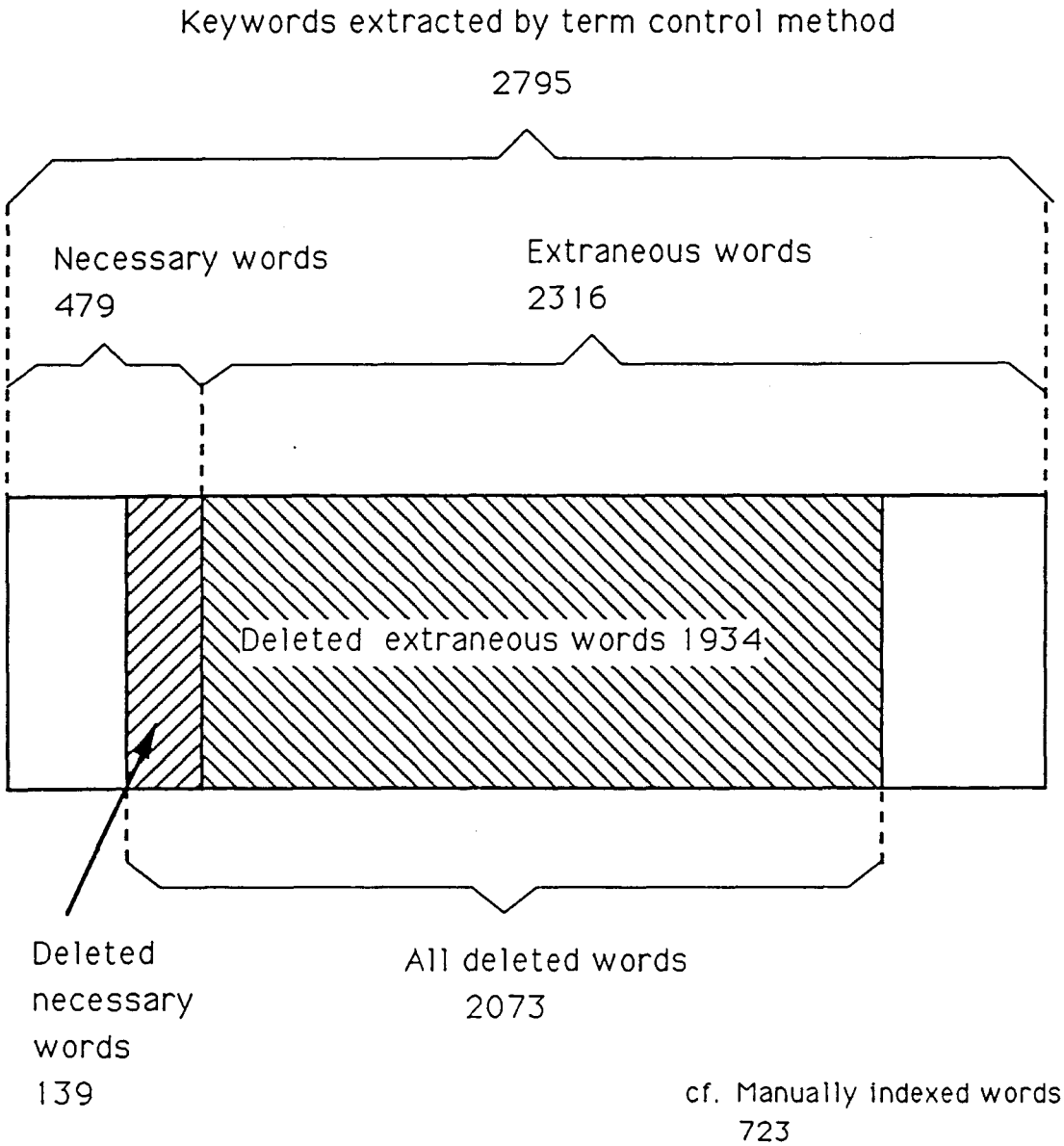



Fig. 6 Keyword selection results.

83.5%. The effect of each deletion algorithm is shown in Fig. 7. From Fig. 7, it would appear that the algorithm using word location information is sufficient for deletion. This algorithm alone, however, does not provide a precise enough rate for commercial use. When the word location algorithm was evaluated separately, the precision rate was 40%--less than the adequate rate, which is over 50%, for commercial use.

### 5.3.2. Keyword selection with rank

The evaluation of the keyword ranking was made using the same sample used for evaluating keyword selection. Again, the evaluation criteria was the percentage of the keywords indexed manually by professional indexers that was replicated by INDEXER. Table 6 shows the results of the evaluation of the ranking function. The results show that nearly 95% of the necessary keywords are included in the top ten keywords. This means that in indexing 10 keywords for newspaper articles, a 95% success rate was achieved using INDEXER's keyword ranking function. This novel keyword ranking technique proved to be very successful in supporting indexing.

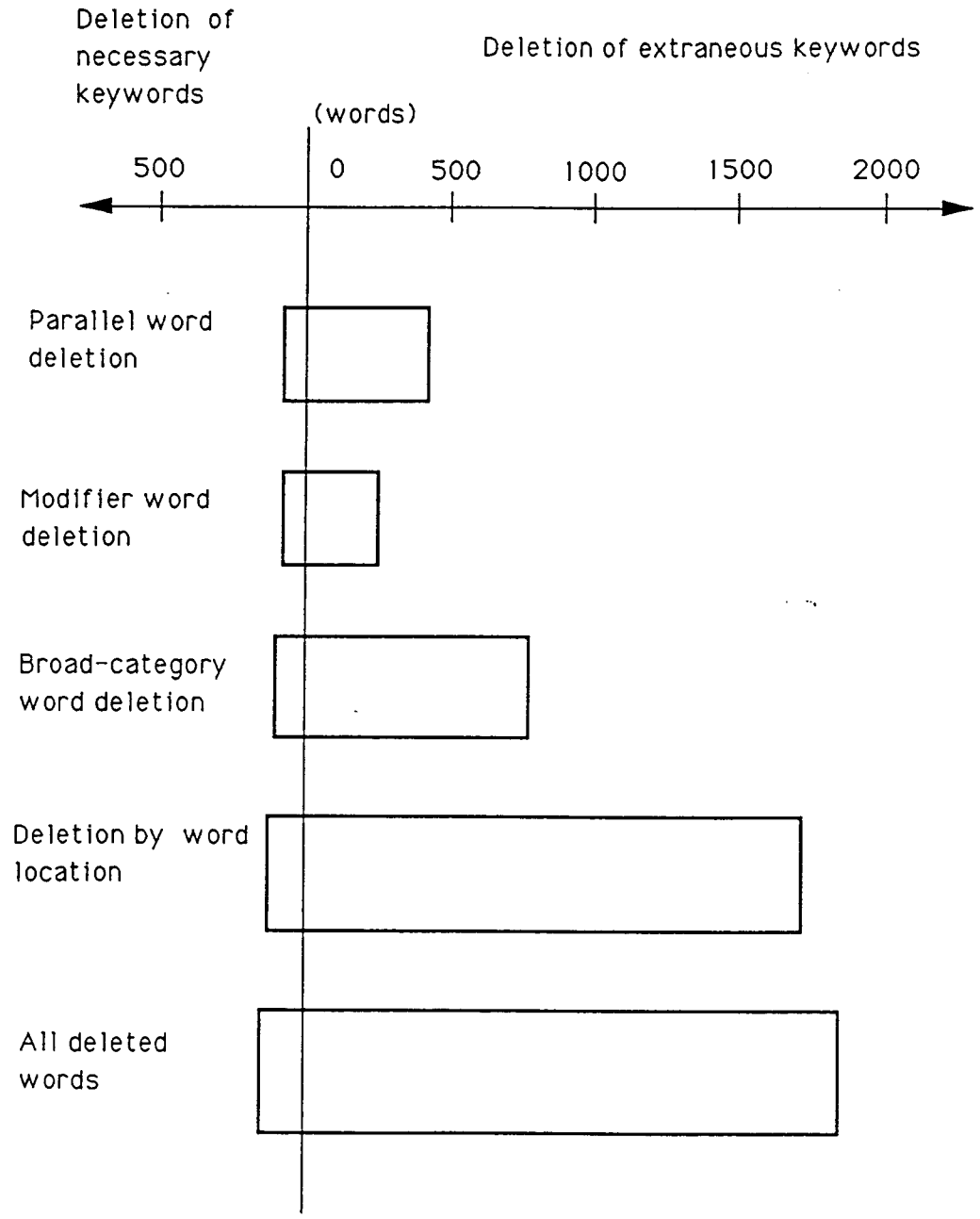



Fig. 7 Effect of deletion algorithms.

Table 6 Evaluation of INDEXER ranking function.

| Ranked keywords | Percentage of necessary keywords included |
|---|---|
| Top 5 | 80% |
| Top 10 | 95% |

# Chapter 6. Applications of INDEXER to information retrieval with association

## 6.1 Information retrieval

Wordprocessors and optical disk filing systems were developed. They became widespread, and it became easy to make machine readable files. As a result, many large-scale databases have been developed. Meanwhile, the development of computer networks enables us to use large-scale databases easily. As a result, many individuals and sections in companies and corporations have began to use them. So, there has arisen a dire need for an individual oriented information retrieval technique as well as a more precise retrieval technique.

The state of the art in the research and development of information retrieval field was described in Chapter 1 (Introduction).

## 6.2 Problems in information retrieval

The problem in information retrieval is the low recall rate and the low precision rate of document retrieval.

Nowadays, only a few AI techniques are used in real-world applications. The conventional AND/OR Boolean search technique is still widely employed but it has the following problems: One is the intricacy of making

AND/OR search queries; another is its inability to retrieve useful documents and the retrieval of useless documents. In other words, the problem is the low recall rate and the low precision rate of document retrieval.

A great number of end users have begun retrieving documents. A new technique is now needed that achieves high preciseness in retrieval, reflecting the end user's intention more explicitly in a more user friendly way.

There are several causes for inaccurate document retrieval. They are the use of keywords that are not indexed to the relevant documents, and the use of keywords that retrieve documents in the too-broad range, or in the too-narrow range compared with the intention of the user. If the user uses keywords that express his intention exactly, there is no problem, but usually, the user can not remember the appropriate keywords. So, there is room for a computer to support the user in remembering the appropriate keywords that express his intention. The new method described in the next section automatically generates these appropriate keywords.

Here, the relationships between query and relevant documents are illustrated in Fig. 8. Case 1 is the use of keywords that are not indexed to the relevant documents. Case 2 is the use of keywords that retrieve documents in the too-broad range. Case 3 is the use of keywords that retrieve documents in the too-narrow range.

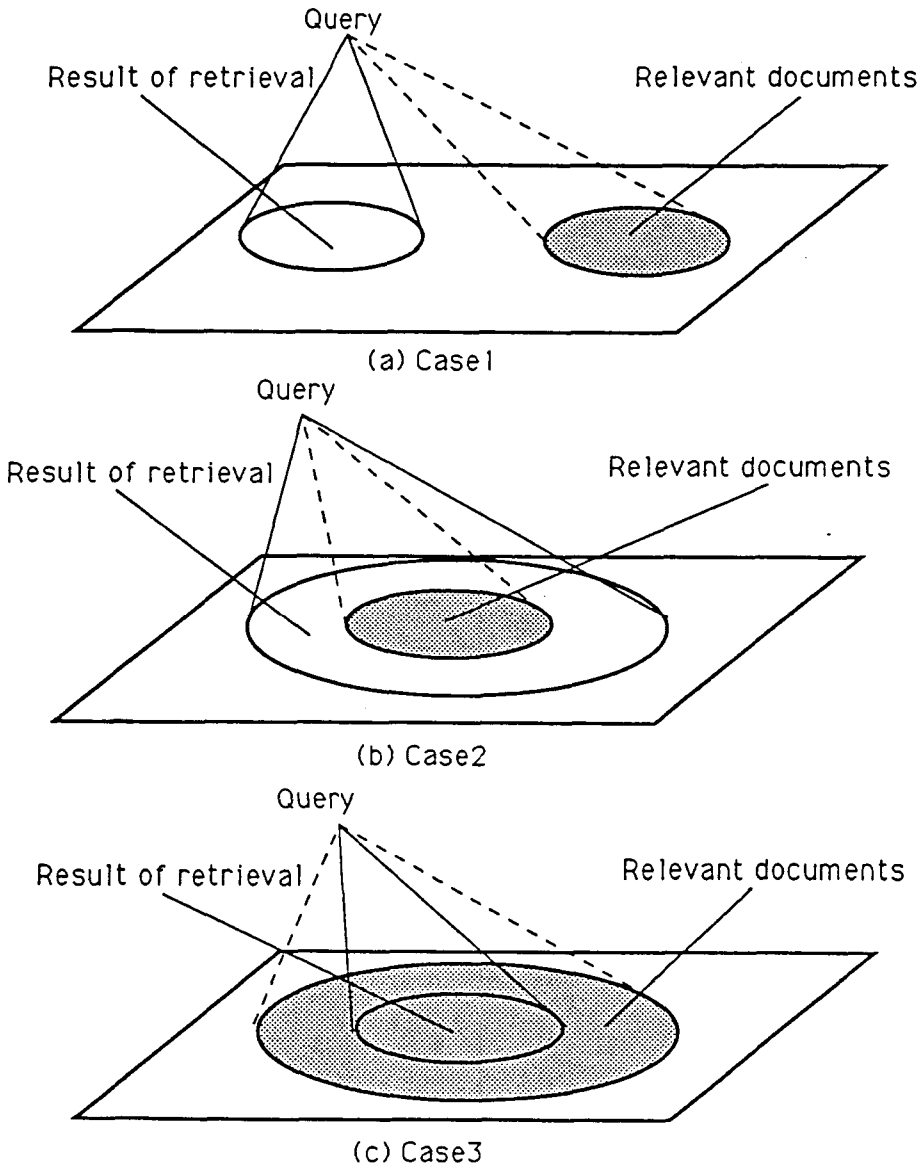



Fig. 8 Relationships between query and relevant documents.

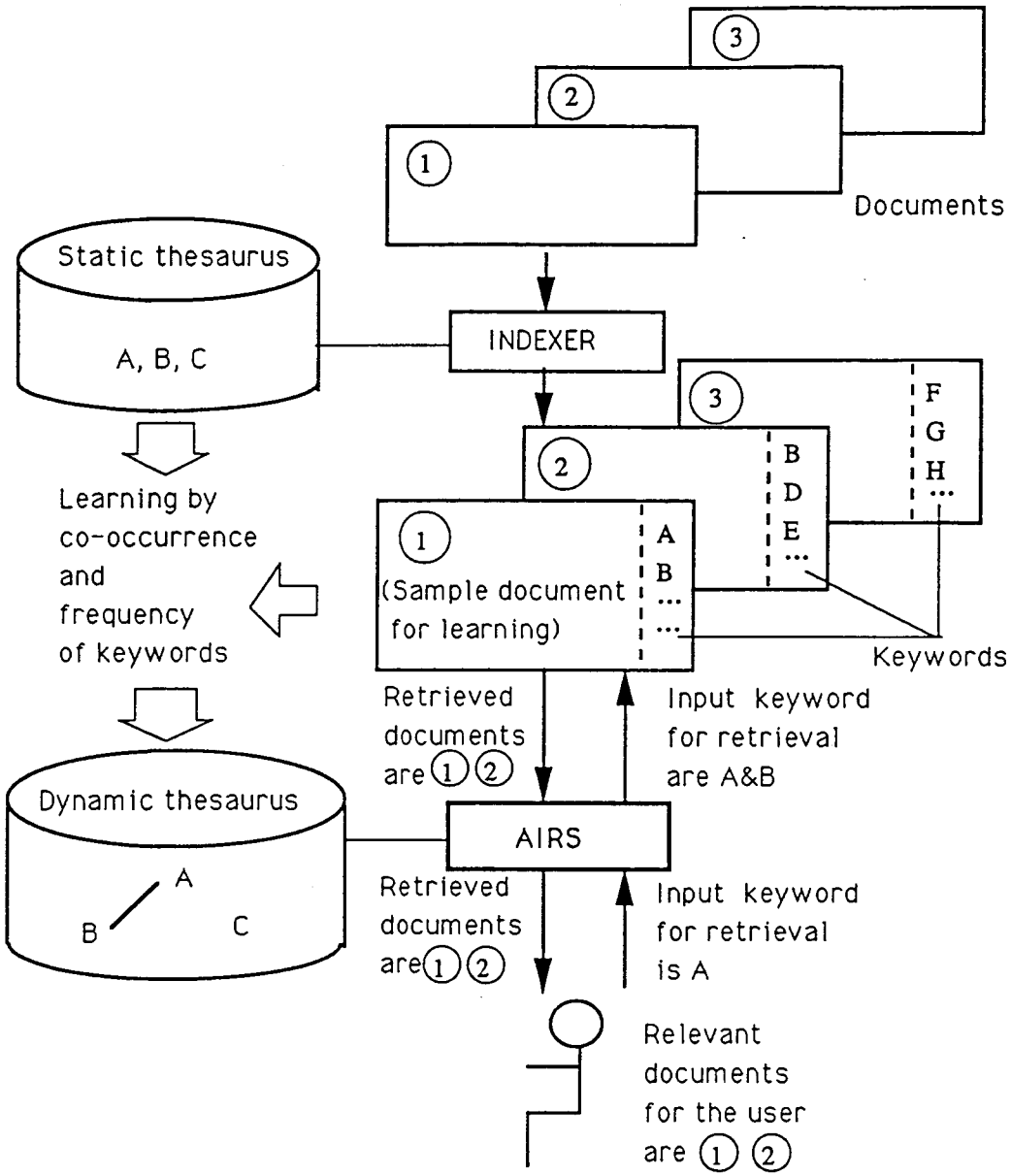
## 6.3 A new method for information retrieval

### 6.3.1 Key idea

A key idea for resolving the problems of the low recall rate and the low precision rate is to use a network of keywords with node weights, which enables generating associated keywords and selecting important keywords. The association of keywords in the network will be useful in generating keywords that are strongly related to the keywords inputted by a user. The node weight of the keywords is useful in selecting keywords that are important to a user. For good utilization of the network, a user should teach the network the linkage of nodes and the weight of nodes that fit his case. INDEXER plays a vital role in making linkages and weighing nodes automatically. A prototype system was designed and implemented using this key idea, and this prototype system was named the associated information retrieval system (AIRS).

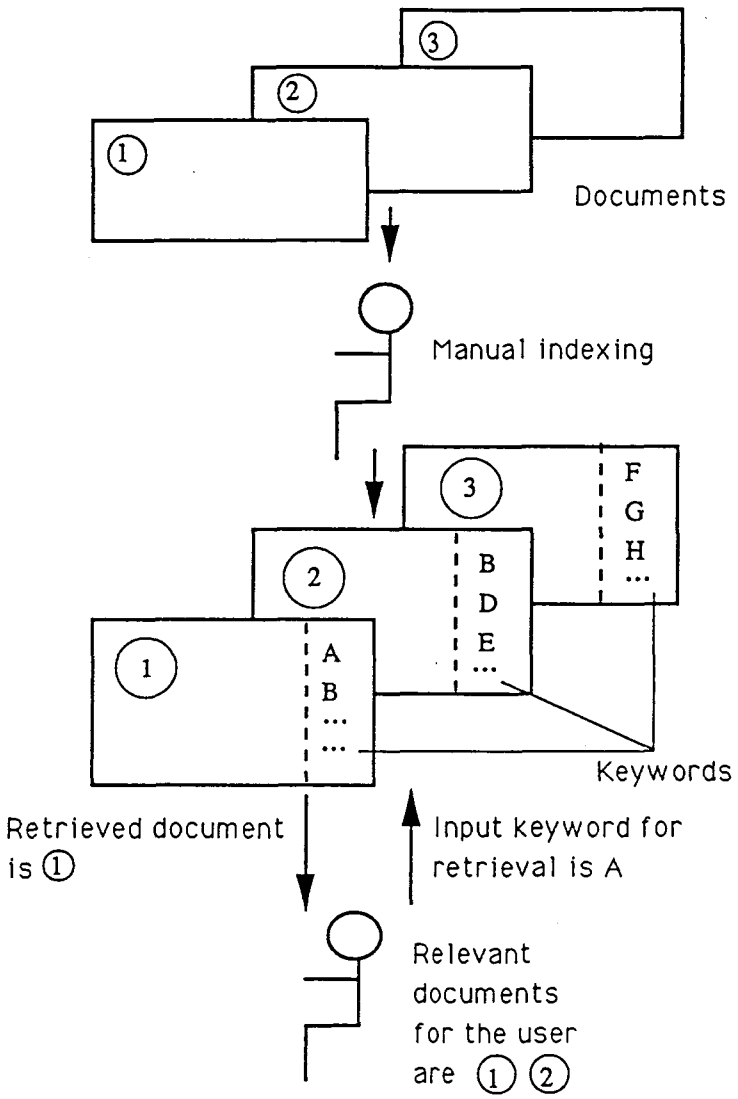
### 6.3.2 Basic concept of the new method

The model of the new information retrieval method is shown in Fig.9(a) as named AIRS Model, together with the conventional model in Fig.9(b). The main points of difference between the conventional model and the AIRS model are as follow. The AIRS model learns a user's interest from sample documents. The user's interest is represented in the dynamic thesaurus using a connection




(a) AIRS model.

Fig. 9 Conventional model and AIRS model.




(b) Conventional model.

Fig. 9 Conventional model and AIRS model.

ist model. Keywords which fit to the user's interest are generated using the dynamic thesaurus, while the conventional model does not generate. In Fig. 9(b), a keyword for retrieval is only A, while in Fig. 9(a), keywords for retrieval are A and B. Keyword B was generated using the AIRS model.

The main idea of this new method is that it extracts a user's interest from the user's sample documents as (a) keywords, (b) the degree of keyword importance in that document, and (c) the linkage between keywords in that document. For extracting keywords and the degree of keyword importance, INDEXER [Kimoto et al. 1989][Kimoto 1991a][Kimoto 1991b] is used. INDEXER is an automatic indexing system for Japanese newspaper articles and technical documents. Keywords are extracted using a thesaurus and they are evaluated by statistical analysis, semantic analysis and grammatical analysis.

The general process flow diagram of AIRS is shown in Fig. 10. Keywords, keyword importance (ranking) and the co-occurrence relation of keywords in a document constitute what we refer to as term information. The dynamic thesaurus is made from the static thesaurus using this term information. The associated keywords are generated from the input keyword of a user and the dynamic thesaurus. These associated keywords are used to retrieve documents that precisely fit a user's own

interest. Extraction and learning a user's interest is done using term information and the dynamic thesaurus, which are described below.

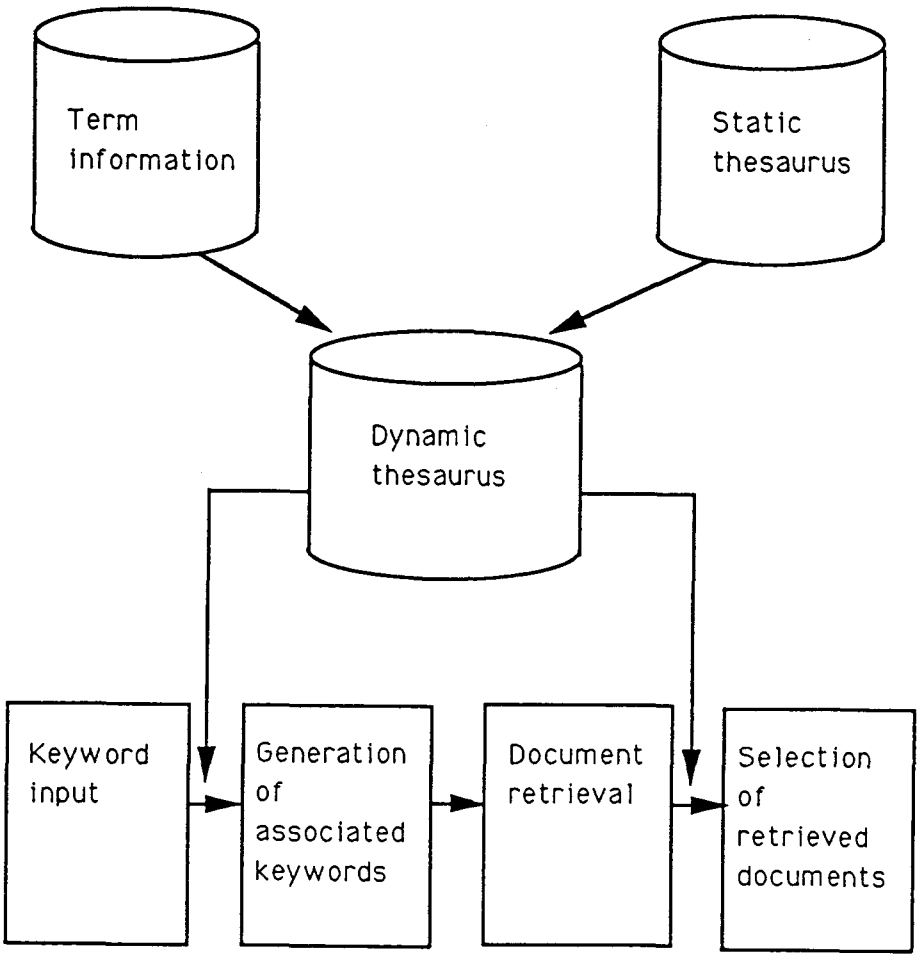
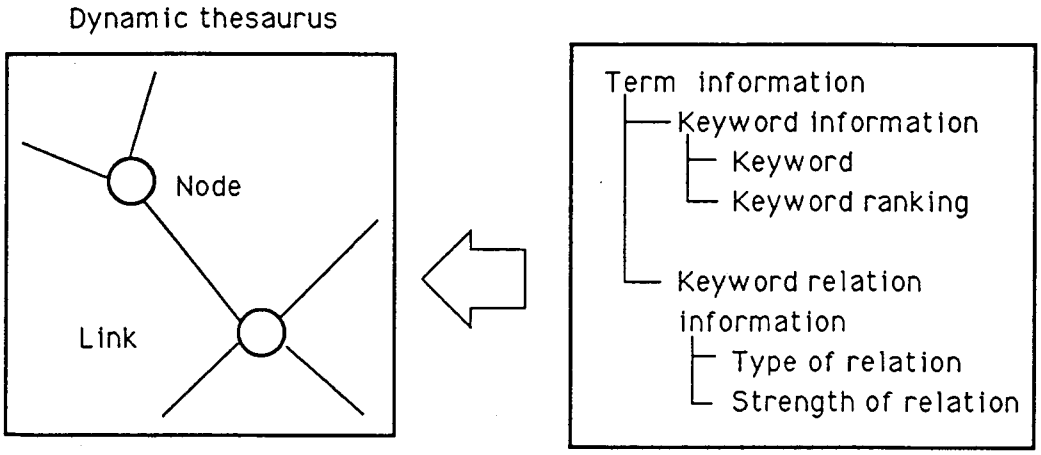



Fig. 10 General process flow diagram of AIRS.

(i) 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. 11(a).

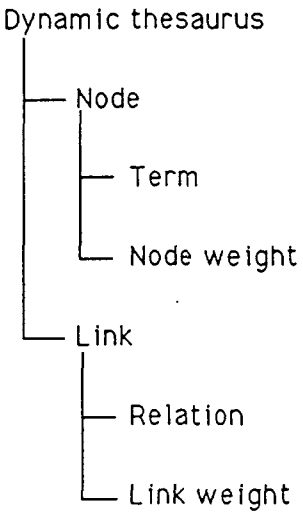



(a) Construction of dynamic thesaurus.

Fig. 11 Construction of dynamic thesaurus using term information.

(ii) 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 the relationship between terms. The data structure of the dynamic thesaurus is shown in Fig. 11(b). Nodes (Term and Node Weight) and Links (Relation




(b) Data structure of dynamic thesaurus.

Fig. 11 Construction of dynamic thesaurus using term information.

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 follow:

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 same sample document (See Fig. 12). This co-occurrence relation is defined very simply. In this definition, a word in an affirmative sentence and a word in a negative sentence are treated in the same way. It may seem that this definition is over-simplified. However, for retrieving documents with the some themes or words, it is important that something be described about the themes or the words in the documents and whether the sentence is affirmative or negative. For example, let us consider two sentences, "Paris is on fire." and "Paris is not on

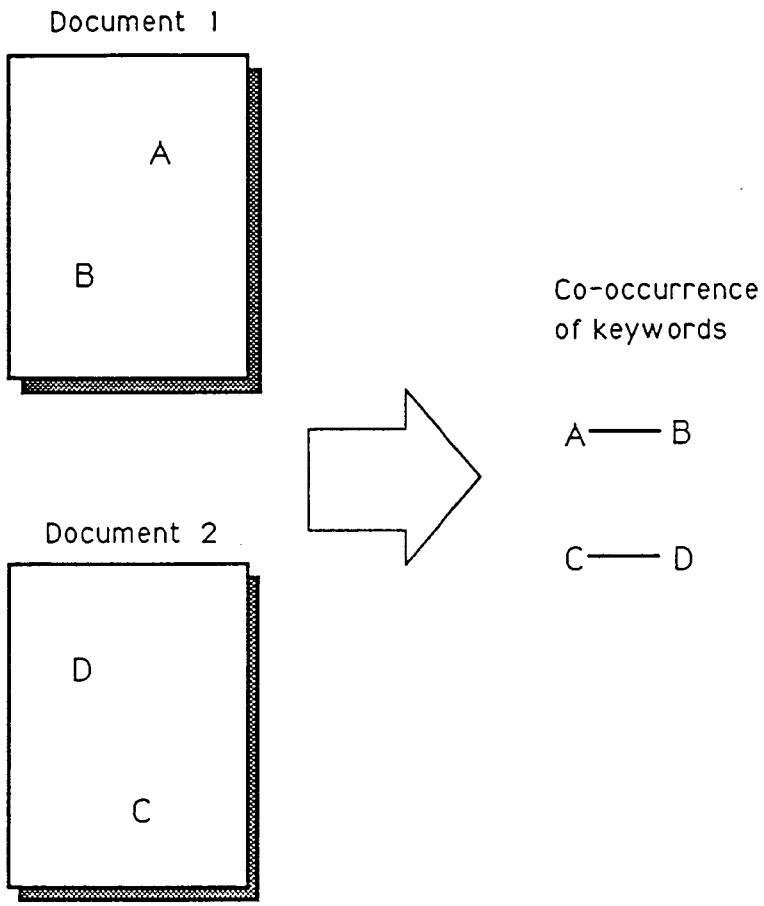



Fig. 12 Co-occurrence of keywords in documents.

fire." These sentences are equivalent in the sense that they describe something about Paris. Most people would be satisfied merely with being able to retrieve those documents whose theme is about Paris. In this sense, therefore, it is not so important to distinguish a negative sentence from an affirmative sentence for retrieving documents.

There are a lot of small, separate networks in the initial state of the dynamic thesaurus (in which state, incidentally, is identical to the static thesaurus). The use of co-occurrence relations and node weights 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.

### 6.3.3 Algorithm

(i) 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. 13).

(ii) 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 INDEXER. It ex-

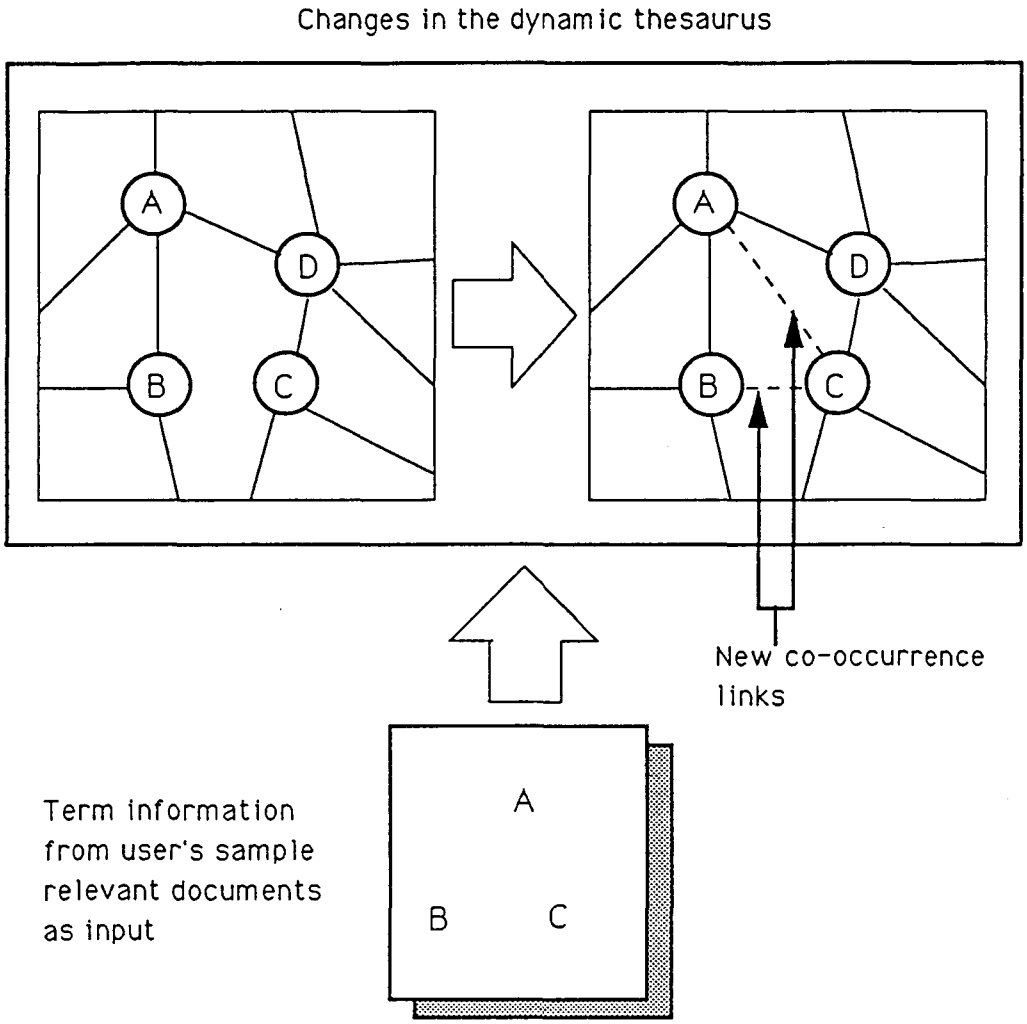



Fig. 13 Generation of links.

tracts and ranks keywords from each of the user's sample documents. Ranking is in the order of importance according to frequency and location in the document. In the formulas presented below, (1)-(4), D is a complete set of sample documents, and Ti denotes individual documents. Ki is a set of keywords extracted from Ti using INDEXER. KWij denotes individual keywords. Assume there are n documents and a total of m keywords in 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) is equal to one (for each i) for normalizing the importance; i.e., 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 each node, denoted as KWr(D) (r=1,...,p, where p is the total number of keywords in D), is calculated as the sum of KI(ij) for each node in D (See Fig. 14).

(iii) Associated keyword generation algorithm

Associated keywords are intended to extend the keywords inputted by a user. Associated keywords are obtained by traversing the links and nodes of the dynamic thesaurus, starting with the node that corresponds to a user's 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 associated keywords. 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,

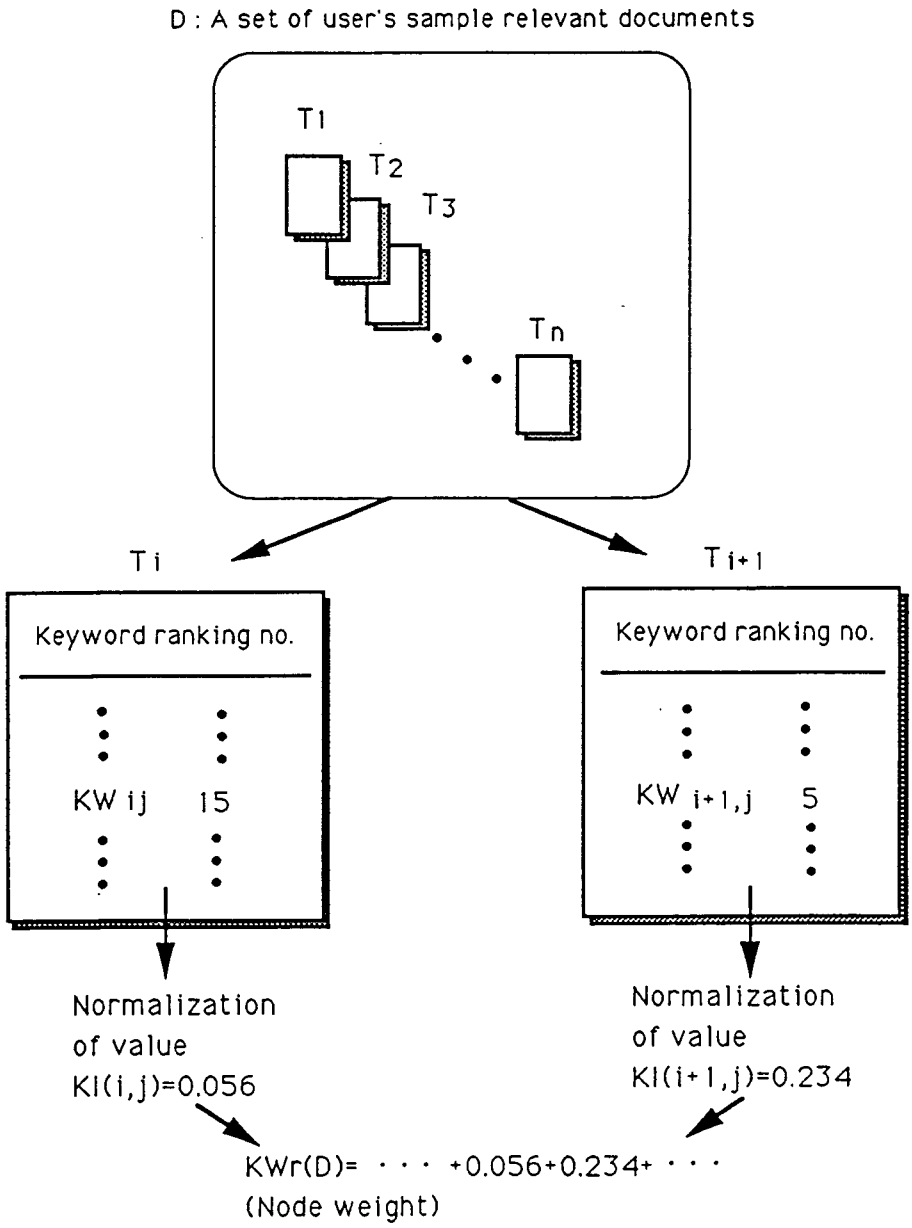



Fig. 14 Node weight calculation.

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 output as associated keywords.

An example of the keyword generation process is shown in Fig. 15. 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. 15. 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.

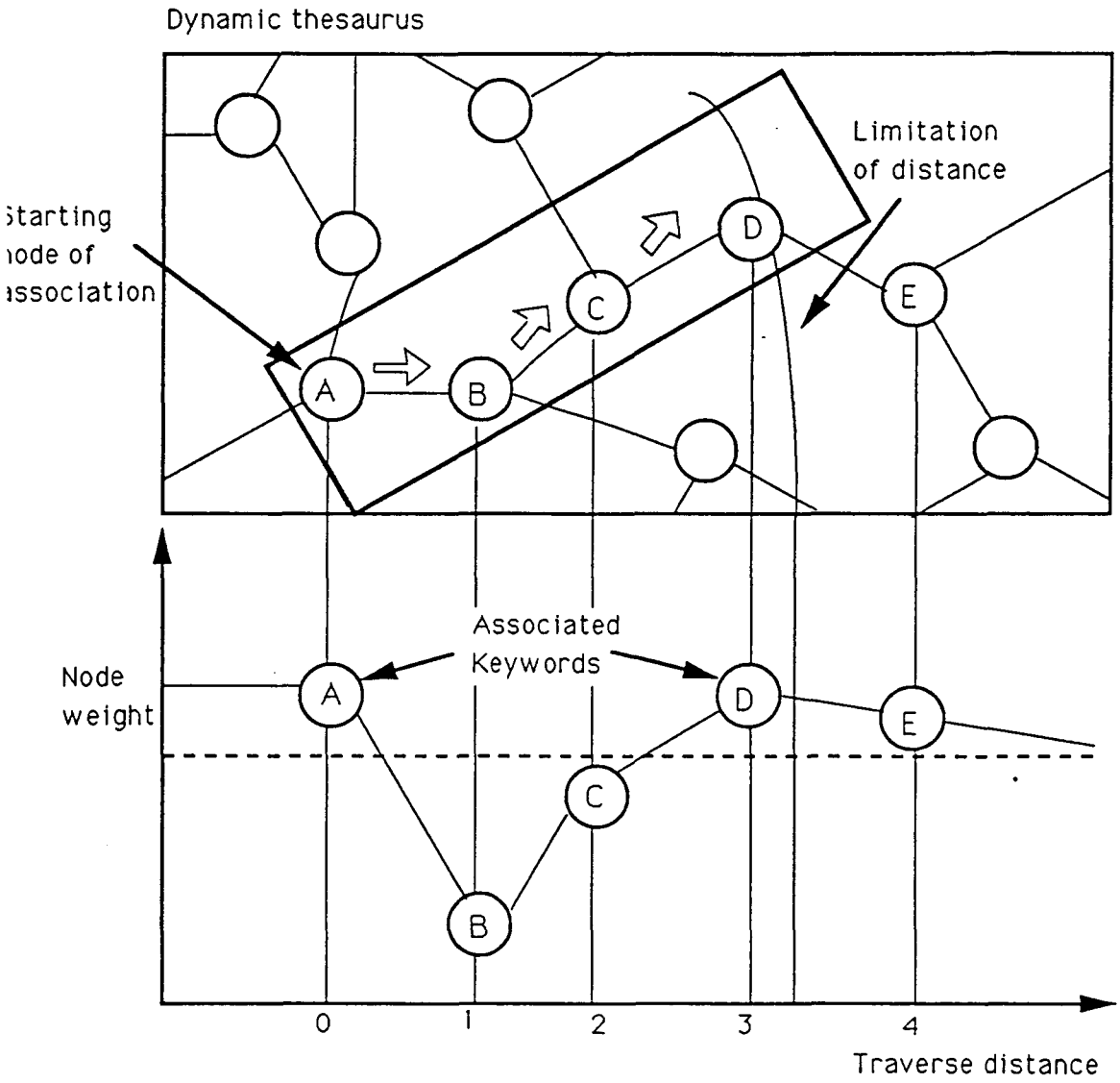



Fig. 15 Generation of associated keywords using links and node weight.

## 6.4 Originality of the new method

The originality of the method used in AIRS is that it determines a user's interests from a 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.

There are several systems in existence similar to AIRS. They are as follows. Miyamoto introduced a fuzzy relation between a pair of keywords for Information Retrieval [Miyamoto et al. 1983][Miyamoto et al. 1986]. Ogawa and T.Morita proposed a learning method using a keyword connection matrix [Morita,T. et al. 1989][Ogawa et al. 1989]. Jung and Raghavan proposed a learning method that gives a positive link weight to semantically similar pairs and a negative link weight to dissimilar pairs by investigating a user's feedback [Jung et al. 1990]. Ito reported the design of a knowledge-based Information Retrieval System that incorporates a knowledge base and an induction mechanism. This knowledge base has a thesaurus whose data structure is a directed graph [Ito 1991]. But in none of these works was a node weight incorporated in the thesaurus or in other data structures. Only Jennings proposed a user model neural network that has both a node weight and a link weight as a data structure [Jennings et al. 1992]. In this neural

network, nodes represent article features (i.e. the first 300 words of an article excluding common words). The nodes are weighed only by the locations of the article features in the article. AIRS has both a node weight and a link as a data structure. In AIRS, the nodes are weighed by statistical analysis, semantic analysis and grammatical analysis of keywords in each sample document. As a result, the nodes are weighed more precisely. As Jennings did not measure the recall rate and the precision rate, which are well-established measurement methods for information retrieval systems [Salton et al. 1983], we can not compare our result with his'. However, our experiments, which will be described in Chapter 7, show that the results obtained with AIRS are much better than those obtained using ordinary keyword search methods.

## 6.5 Other applications of INDEXER

In this section, other applications of INDEXER are described.

There are two kinds of databases used in printing companies. One is text databases such as books and dictionaries and the other is Chinese character databases such as name lists, book indexes and catalogues. There are two ways of utilizing these databases. One way is to use them for making paper materials and the other is to use them for making electronic media such as CD-

ROMs. In each case, an original database is made and then data are compiled to make paper materials or electronic media. INDEXER is used in compiling these databases.

### 6.5.1 Making paper materials

(i) Book index compilation support

In most types of books, it is usually necessary to compile an index. INDEXER aids in doing this work. Before INDEXER was developed, a compiler had to mark the keywords in a volume. The marked keywords were then extracted by a computer system, and a compiler used them to compile the index. In marking the keywords throughout a volume, a compiler must always be careful not to omit any of them. However, since a book index compiler is a human being, some omissions are inevitable. When INDEXER is used in doing this work, the compiler first gives the system a list of keywords. INDEXER then marks and extracts these keywords from the volume automatically. After this process is completed, the compiler checks the results. As a result, it is not necessary for the compiler to mark the keywords in a volume.

(ii) Providing rubies along with Chinese characters

It is often necessary to provide readings along with Chinese characters in works of literature, history books, textbooks, etc. in accordance with the ages of

the readers and the fields being written about. These readings supplements are referred to as "rubies" by printing companies. Since the primary school and junior high school students can read only a comparatively small number of Chinese characters, it is necessary to provide readings along with the Chinese characters, either wholly or partially, in books for these students. It is also necessary to provide readings along with the Chinese characters in history books, literary classics, scientific textbooks, etc., because they contain Chinese characters which are used so rarely or which are so old that most peoples cannot read them. Accordingly, INDEXER, which uses NLP technology to automatically provide readings, is used for that. The inputs to INDEXER are texts consisting of all the different kinds of characters used in Japanese writing, i.e. "Chinese characters", "katakana" and "hiragana". The texts are analyzed using a morphological analysis program in INDEXER and readings are assigned to Chinese characters through a function included in INDEXER that provides the readings of these characters (See Fig. 16). After this process is completed, text data with readings are inputted to the Computer Typeset System (CTS) for making materials.

【原文】

病原体のウイルスのように、コンピュータからコンピュータに”伝染”し、利用者のソフトを破壊していくプログラムのこと。一九八七年ごろから、米国のハッカー（→別項）の間で、パソコン通信を媒介とした「トロイの木馬」とよぶ破壊プログラムが流行し始め、これがさらに進化し、悪質になったのが、「コンピュータ・ウイルス」で、この方は侵入後に自己増殖していくので始末が悪い。ＩＢＭの地域ネットワークに侵入したり、イスラエルのヘブライ大学コンピュータ・センターの貴重な資料をダメにしたり、その被害は拡がっている。コンピュータ・セキュリティ（防護）の専門家たちが”免疫プログラム”や”抗体プログラム”を考察中。

【総ルビカナ】

〈病原　ビョウゲン〉〈体　タイ〉のウイルスのように、コンピュータからコンピュータに”〈伝染　デンセン〉”し、〈利用　リヨウ〉〈者　シャ〉のソフトを〈破壊　ハカイ〉していくプログラムのこと。〈一九八七　センキューヒャクハチジューナナ〉〈年　ネン〉ごろから、〈米国　ベイコク〉のハッカー（→〈別項　ベッコウ〉）の〈間　アイダ〉で、パソコン〈通信　ツウシン〉を〈媒介　バイカイ〉とした「トロイの〈木馬　モクバ〉」とよぶ〈破壊　ハカイ〉プログラムが〈流行　リュウコウ〉し〈始　ハジ〉め、これがさらに〈進化　シンカ〉し、〈悪質　アクシツ〉になったのが、「コンピュータ・ウイルス」で、この〈方　カタ〉は〈侵入　シンニュウ〉〈後　ゴ〉に〈自己　ジコ〉〈増殖　ゾウショク〉していくので〈始末　シマツ〉が〈悪　ワル〉い。ＩＢＭの〈地域　チイキ〉ネットワークに〈侵入　シンニュウ〉したり、イスラエルのヘブライ〈大学　ダイガク〉コンピュータ・センターの〈貴重　キチョウ〉な〈資料　シリョウ〉をダメにしたり、その〈被害　ヒガイ〉は〈拡　ヒロ〉がっている。コンピュータ・セキュリティ（〈防護　ボウゴ〉）の〈専門家　センモンカ〉たちが”〈免疫　メンエキ〉プログラム”や”〈抗体　コウタイ〉プログラム”を〈考案　コウアン〉〈中　チュウ〉。

Fig.16　Example of giving readings along Chinese characters.

### 6.5.2 Making electronic media

INDEXER is also used for generating indexes during the process of making CD-ROMs. These indexes are used for looking up data in CD-ROMs. There are two methods of extracting keywords from a database to make an index. One is free keyword extraction, and the other is control keyword extraction. In making the index for a CD-ROM, INDEXER is used to freely extract keywords from the databases that are to be stored in CD-ROMs. To look up words stored in dictionaries in CD-ROMs, it is necessary to be able to use many keywords so as to look up entry words of a dictionary. Therefore, all keywords are freely extracted from the text of each entry word in the dictionary. And when the freely extracted keywords are compound nouns, all components of these compound nouns are treated as keywords. For example, "natural", "language", "processing", "natural language" and "language processing" are generated as new keywords from "natural language processing" and are used as indexes.

### 6.5.3 Making Chinese character databases

Chinese character databases contain only column type data. Examples of Chinese character databases include name list and catalogue databases. Name list databases consist of columns for "name", "address", "affiliation", "place of birth", "alma mater" and so on.

Catalogue databases consist of columns for "name of book", "author's name", "publisher", and so on. The contents of these columns are written in Chinese characters, and the databases are maintained and manipulated using each column. Data are sorted and classified frequently using these columns in order to meet various user demands. The readings of the Chinese characters are used as the key for sorting or classification. NLP technology is used to provide readings along with Chinese characters in each column of the database, and new columns, such as readings of names or addresses, are automatically generated.

### 6.5.4 Indexing a very large-scale text database

This section describes an application of INDEXER to a very large-scale text database. The database includes many different kinds of articles taken from 37 Japanese newspapers and about 120 magazines. The 37 newspapers include daily newspapers, trade papers and other professional papers, and others. The 120 magazines include amusement magazines, magazines for daily life, magazines for professional engineers and others. The database contains 2.5 million articles, and about 2000 articles are added daily. The applied indexing method is free term indexing and terms are weighed according to their importance in each article. For indexing the very large-scale text database, a very large-scale dictionary file,

which contains words from almost all fields, was made. This dictionary contains about 930,000 items. Since the weights of each algorithm are open to users of INDEXER, the user can give the best weight set to INDEXER; the one which best fits their own indexing scheme. This weight set is called the "user's weight table." The number of keywords for an article is proportional to the length of that article. An evaluation of the indexing showed that the keyword recall rate was 70.9% and the keyword precision rate was 30.7%. For this evaluation, 114 articles from two newspapers were used, and about 500 articles from 24 newspapers and magazines will be used in the next evaluation. The results of the first evaluation were satisfactory for the user.

# Chapter 7. Evaluation of AIRS

## 7.1 The AIRS system

This section proposes a new system that incorporates a 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. The system was designed on the idea of a new information retrieval method described in Chapter 6. This new system is called the Associated Information Retrieval System (AIRS).

AIRS is a prototype system. A process schematic of AIRS is shown in Fig. 17. The AIRS operates as follows.

Step 1. Keywords and the keyword ranking are extracted from the sample relevant document using INDEXER.

Step 2. Term information is constituted from keywords, keyword ranking, and keyword co-occurrence relation in a document.

Step 3. The static thesaurus is modified by term information to form a dynamic thesaurus. Links are generated and node weights are calculated while the dynamic thesaurus is being made.

Step 4. 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

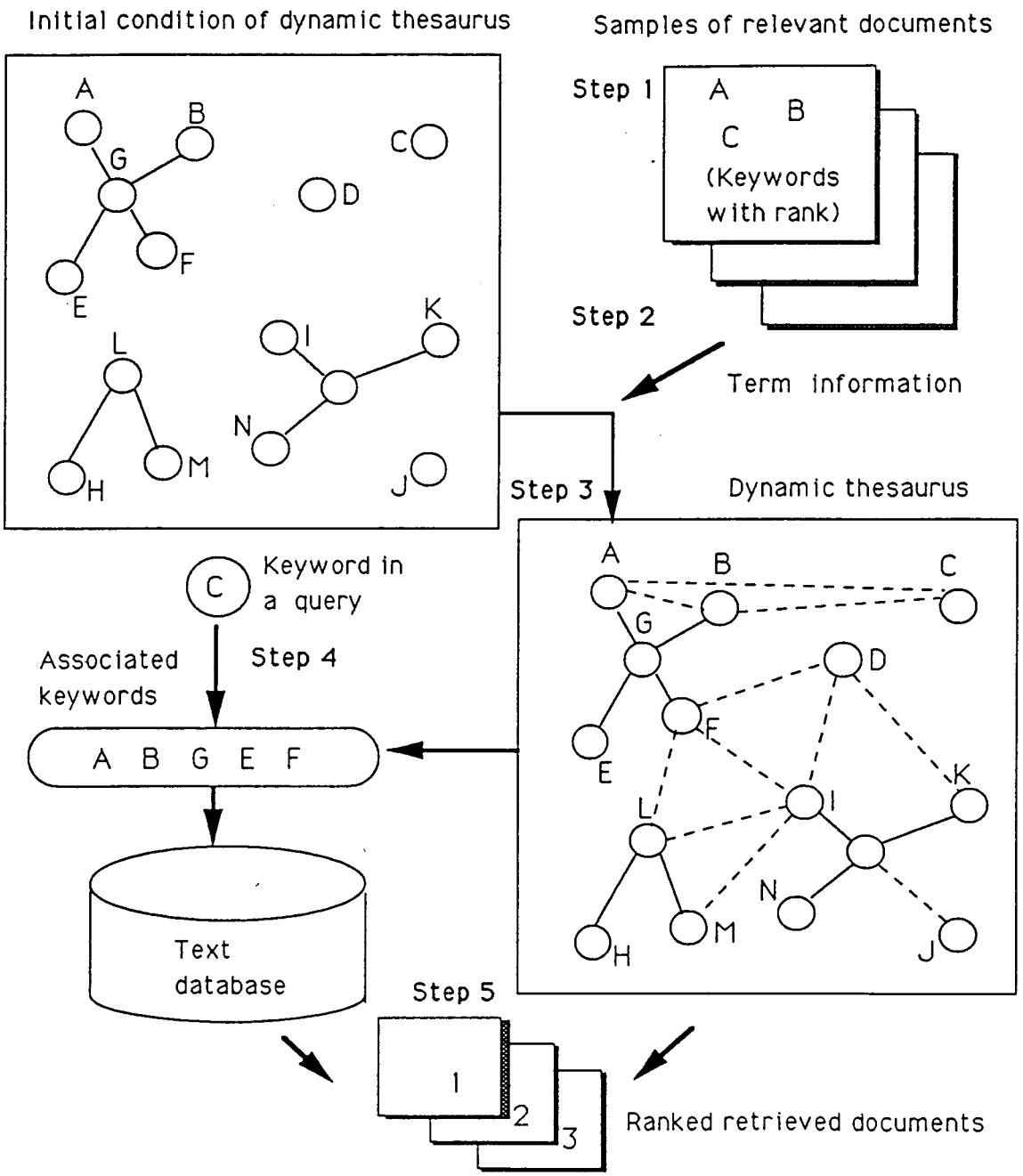



Fig. 17 Schematic of associated information retrieval.

their node weights and links.

Step 5. Documents are retrieved using these associated keywords. Retrieved documents are ranked by using information in the dynamic thesaurus.

## 7.2 Experimental data

A prototype of AIRS was implemented, and an experiment was conducted using this system. The results of this experiment are described in the next two sections, with the results of the construction of the dynamic thesaurus described in Section 7.3.2 and the results of the document retrieval described in Section 7.3.3.

The experiment was carried out using a database of 163 Japanese newspaper articles. The average number of Japanese characters in each article was 500. A thesaurus that had already been made for retrieving newspaper articles was used as a static thesaurus. This static thesaurus has about 9,000 terms. All five of the different kinds of links (described as relations between nodes in Section 6.3.2(ii)) are traversed in order to generate associated keywords.

## 7.3 Evaluation

### 7.3.1 Evaluation criteria

The accuracy of document retrieval is measured in terms of the document recall rate and the precision

rate. Both a high recall rate and a high precision rate are necessary for accurate and effective document retrieval. These rates are defined as follows:

$$Dr = \frac{Na}{Nrel}, \quad \text{and}$$

$$Dp = \frac{Na}{Nret},$$

Dr is the document recall rate and Dp is the precision rate, where Na is the number of the relevant documents that are duplicated by the retrieved documents, Nrel is the number of the relevant documents and Nret is the number of retrieved documents.

### 7.3.2 Construction of dynamic thesaurus

The weights of the nodes were changed and new links were made between the nodes while the dynamic thesaurus was being made. After constructing the thesaurus, the weights and the new links were checked to see if they were correctly made.

In checking the correctness of the linking, it was very difficult to decide if the nodes (terms) had a linkage to each other or not. For example, when the sentence "The tennis championship tournament was held in Paris." is in a sample document, a new link is made between the keywords "tennis championship tournament"

and "Paris". Is this linkage correct or incorrect? Usually there is no linkage between these two keywords. In this particular case, however, they have a strong linkage. Therefore, we made the assumption that all of the keywords in a sample document have a linkage to each other. It is undoubtedly true that there are both strong and weak linkages, and several different kinds of links, such as location-links, agent-links and so on. However, these have not been incorporated into AIRS yet.

In order to check whether the nodes were weighed in such a way that they represent the user's interest correctly, the weights of each node were checked by users after the dynamic thesaurus was constructed. In the checking process, experiments were conducted eight times, using different sample documents and checking was done for each experiment. Here, let us examine the checking method and the result of one of the eight experiments. The dynamic thesaurus has about 9,000 nodes, and the weights of all nodes are 0.0 initially. The dynamic thesaurus was constructed using four sample documents in this experiment. After the construction, the weights of 55 nodes were increased. These 55 nodes were divided into three groups, A, B and C, by the user. Nodes in A group have a strong relevance to the user's interests, which are in his mind and which were expressed by the sample documents. Nodes in B group have a weak relevance to the user's interest, and nodes in C

group have no relevance to the user's interest. The number of nodes in groups A, B and C is 22, 8 and 25, respectively. If the nodes in A group have a larger node weight than the nodes in B group, and the nodes in B group have a larger node weight than the nodes in C group, the dynamic thesaurus represents the user's interest well, otherwise it does not. In Fig. 18 the

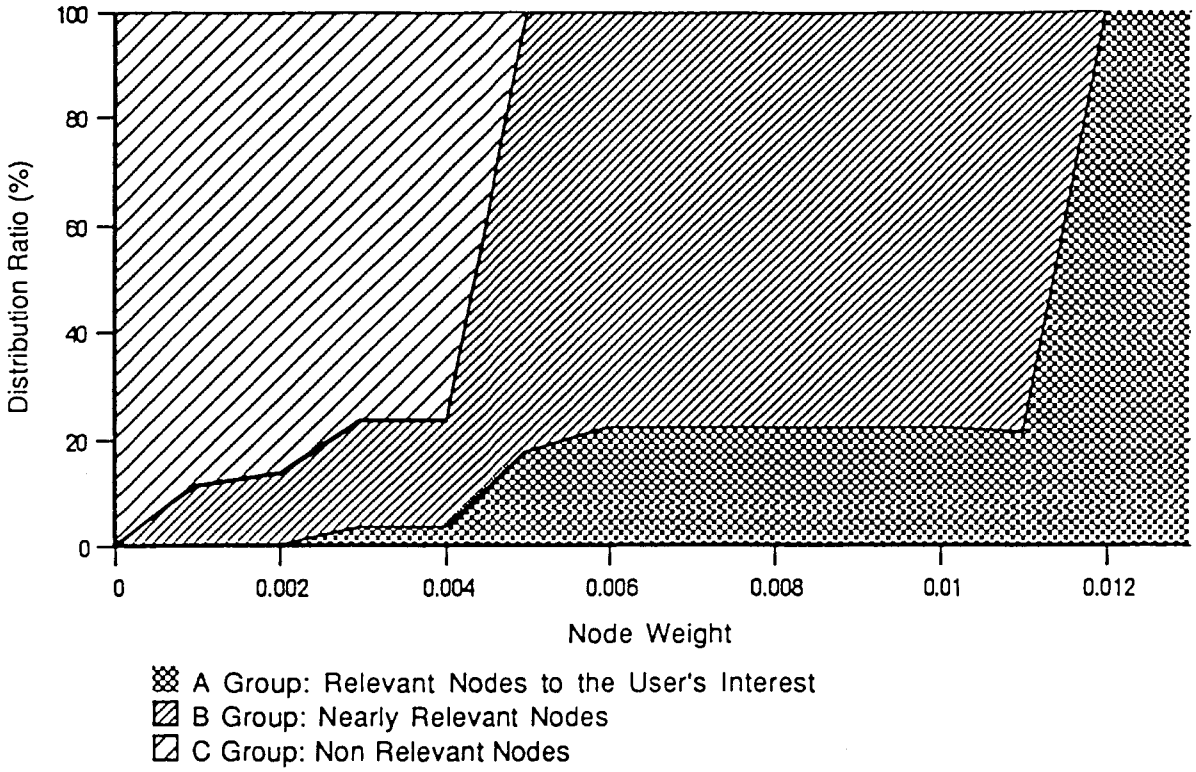



Fig. 18 Distribution of nodes in dynamic thesaurus.

distribution ratio between each group is shown for various node weights. The result shows that, generally, the A group nodes have larger node weights than the B group nodes and the B group nodes have larger node weights than the C group nodes. In the other seven experiments, the same results were obtained. Thus, the nodes were correctly weighed and the dynamic thesaurus represented the user's interest well.

### 7.3.3 Document retrieval using AIRS

(i) Effects of term information on document retrieval

Document retrieval experiments were performed using the associated keywords generated by AIRS. In this section, the effect of term information on document retrieval is described.

Any user of AIRS can make term information the way as he wants. The contents of term information depend on the sample documents that a user selects for making term information. There are several ways of selecting sample documents. They are as follows:

A: A complete set of relevant documents is selected.
B: Half of the relevant documents are selected.
C: Semi-relevant documents are selected.
D: Non-relevant documents are selected.
E: No documents are selected.

Chapter 7

Newspaper articles were chosen as sample documents. The experiments were conducted twice. In each experiment, different sample relevant documents were selected and different keywords for retrieval were input and different documents were retrieved. The numbers of sample documents used to extract term information were;

A:4, B:2, C:2, D:2, E:0, for the first experiment.
A:4, B:2, C:2, D:2, E:0, for the second experiment.

By using these five types of documents for each experiment, five types of term information were extracted, and document retrieval was then carried out for each type. The Boolean OR search strategy was adopted as the search strategy in retrieving the documents. Both the recall rate (Dr) and the precision rate (Dp) were evaluated.

The results of the experiments are shown in Fig. 19. As can be seen, both Dr and Dp increased in the order of D, E, C, A, B for the first experiment and D, E, C, B, A for the second one, with D being documents quite different from the relevant documents and A being the relevant documents themselves.

For both A and B, the Dr and Dp increased more than they did for E. This is because node weights had been increased and many links had already been made between

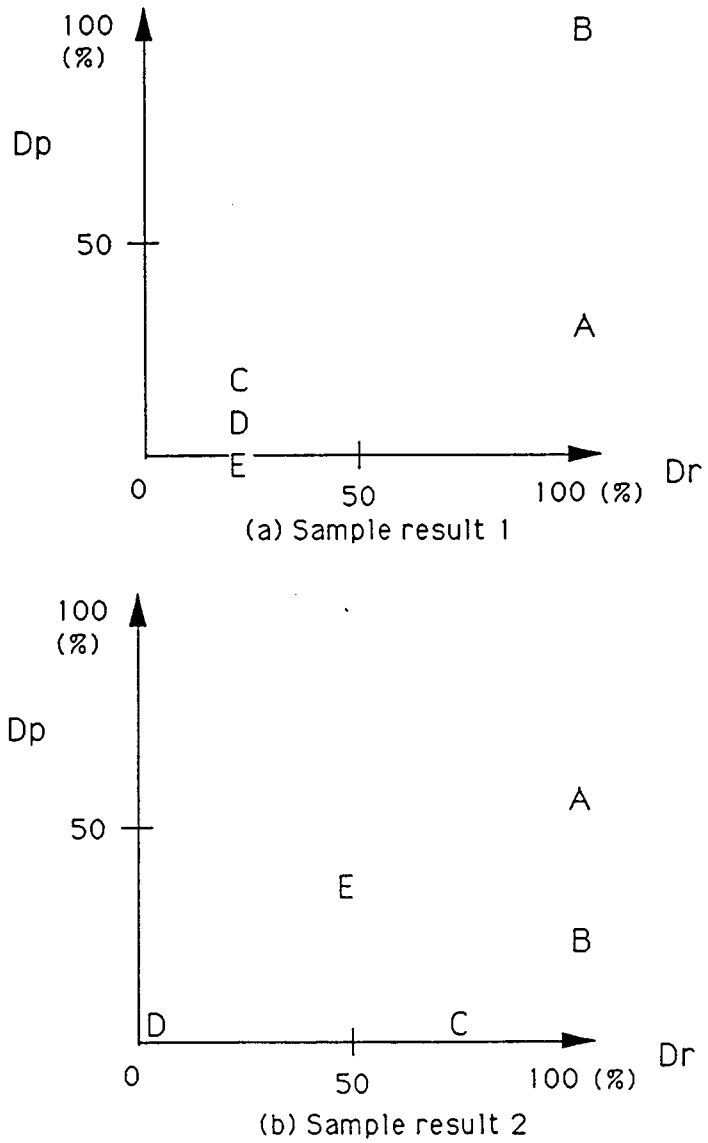



Fig. 19 Sample documents and corresponding results of document retrieval.

nodes around the user-input keyword in the dynamic thesaurus using the sample relevant documents. As a result, associated keywords, which coincided with the user's search intention, were generated and relevant documents were retrieved using those associated keywords.

On the other hand, for D, both the Dr and Dp decreased more than they did for E. This is because links had been made and the node weight had been increased using non-relevant documents in an area that was different from the user's area of interest.

Eight more experiments were conducted with a database containing 800 documents. These were carried out in the same way as the two experiments described above, and the same results were obtained.

It can be safely concluded that associated keyword generation reflects well the contents of the term information, and that by using these associated keywords, accurate and effective document retrieval becomes possible. In other words, if the user gives more appropriate documents to AIRS, he can get more accurate results, and contrariwise, if he does not, he can not. This means that the mechanism of AIRS simulates the human way of incorporating a user's interest in a computer memory.

(ii) Effects of threshold value on Dp and Dr

This section describes a way of determining a

threshold value that indicates good retrieval results. An experiment was conducted in order to find a threshold value that indicates good Dr and Dp, and the results are shown in Fig. 20. From these results, we can readily see that as the threshold value increases, the Dr decreases and Dp increases. We repeated this experiment more than 30 times, changing the sample relevant documents and keywords for document search each time. In every experiment, we got the same pattern of results shown in Fig. 20. These results indicate that the best threshold value lies between 70 and 80% of the largest node weight of the associated keyword. In reality, however, the number of documents retrieved using a threshold value smaller than the best one was only one-tenth of the number of documents retrieved using the threshold value of 0. This is quite a good document retrieval result, and is shown in Fig. 20 as the curve labeled "number of retrieved documents". All of these experiments were conducted for a user having a single interest. Another experiment will be necessary for a user having multiple interests.

In AIRS, the same threshold value is applied regardless of the traversing distance for the associated keyword generation. Both a higher Dr and a higher Dp could be achieved by introducing a variable threshold value that is dependent on the traversing distance.

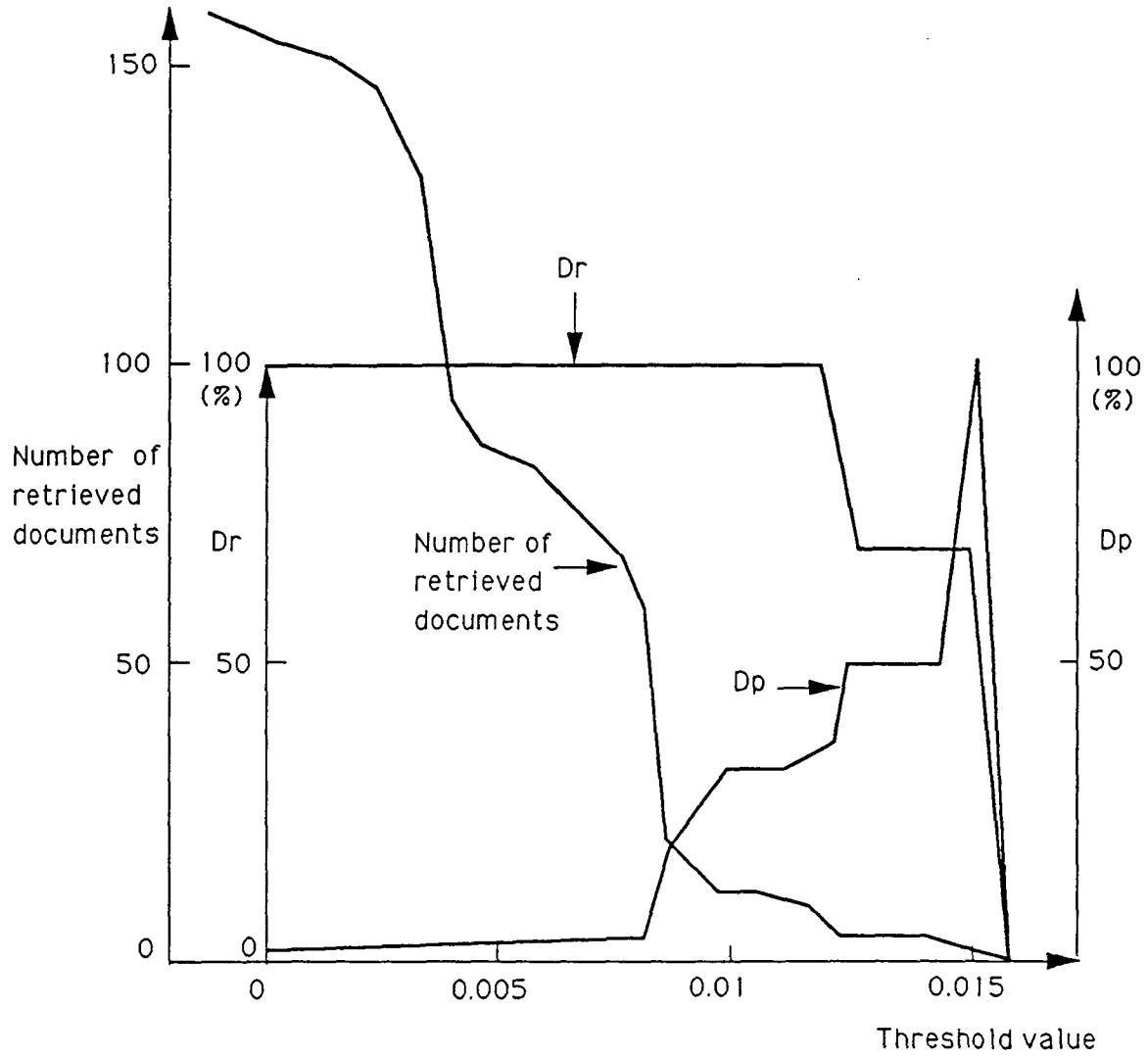



Fig. 20 Threshold value and results of document retrieval.

(iii) The relationship between traversing distance for associated keyword generation and document retrieval

An increase in the traversing distance for associated keyword generation brings about the generation of many associated keywords, including a few relevant keywords (nodes) and a lot of non-relevant keywords (nodes). Document retrieval using these associated keywords results in a high Dr and a low Dp. The threshold value is effective in deleting the non-relevant keywords from the associated keywords. An experiment on document retrieval involving a change in the traversing distance was conducted and the result of the experiment is shown in Fig. 21. In this experiment, the threshold value was set at 0.012. Of course, the result varied according to the threshold value. It would be possible to obtain a better Dr and Dp by introducing a variable threshold value, as was described in the paragraph just before.

(iv) Efficiency of AIRS

In order to evaluate the efficiency of AIRS, an experiment was conducted on a VAX 8800 mini-computer using a document database of 800 newspaper articles, each containing approximately 500 Japanese characters. An 800-document database is fairly practical for use as a personal database. The results of the evaluation are

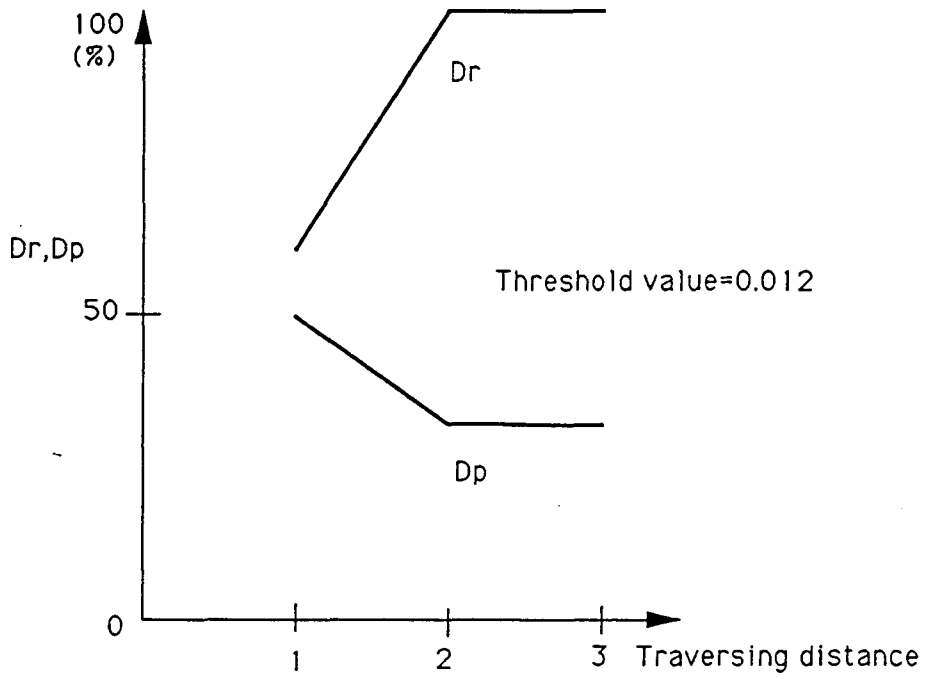



Fig. 21 Result of document retrieval.

depicted in Table 7. The time needed to construct the dynamic thesaurus from the static thesaurus, using a sample relevant document, was 60 seconds. The used sample relevant document was a newspaper article, and the number of Japanese characters in it was 956, and the number of keywords extracted from it was 15. The associated keyword generation time and the document retrieval time are listed for each traversing distance for associated keyword generation. As AIRS is now a prototype system, it has some time-consuming processes, such

Table 7 Efficiency of AIRS

| Traversing distance | Associated keywords generation time | Document retrieval time |
|---|---|---|
| 1 | 14 (sec) | 14 (sec) |
| 2 | 21 | 30 |
| 3 | 41 | 35 |

as the need to use the file open command whenever accessing the thesaurus file. The dynamic thesaurus constructing time, the associated keyword generation time and the document retrieval time will be considerably shorter in the next version of AIRS since these time-consuming processes will be eliminated.

## 7.4 Other expected effects of AIRS

The expected effects of AIRS are as follows:

(i) Associated keywords that reflect the user's interest are generated by the dynamic thesaurus.

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

(iii) 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 a user's prior interests would be possible.

# Chapter 8. Conclusions

## 8.1 Summary

### 8.1.1 Automatic indexing

The free term method, widely used for indexing newspaper articles, has the disadvantage of extracting a great many extraneous keywords. The proposed method of automatic selection and ranking of keywords uses linguistic analysis, experts' indexer knowledge, and word location and/or word frequency in the text to select keywords. A system implemented using this method extracts keywords from newspaper articles, and ranks them in the order of importance.

As for the keyword selection function, the experiment showed that this system achieves 50% for both a recall rate and a precision rate, compared with a recall rate of 70% and a precision rate of 10% for the existing free term method. For a recall rate, 50% is high enough for use. This level of precision is much closer to the 70 to 80% of manual indexing. A precision rate of 50% means that one out of two keywords extracted by the system is significant to the newspaper article.

As for the keyword ranking function, the results showed that nearly 95% of the necessary keywords were included in the top ten keywords. This means that in indexing 10 keywords for newspaper articles, a 95% success rate was achieved using INDEXER's keyword rank-

ing function.

These novel keyword selection and ranking techniques proved to be very successful in supporting indexing. This new system holds considerable promise for summarizing, classifying, understanding, and extracting knowledge from text.

### 8.1.2 Information retrieval with association

A new method of information retrieval was described. This method introduced a dynamic thesaurus that consists of nodes and links. Each node and link corresponds to a term (keyword) and a relation of terms. The dynamic thesaurus incorporates the user's interest in retrieving documents by changing the node weight and making new links between terms using the user's sample relevant documents. A document retrieval experiment was carried out using the dynamic thesaurus and both a high-recall and a high-precision rate were attained. This leads to the conclusion that the dynamic thesaurus is effective for highly precise document retrieval and that the retrieved documents fit the user's interest very well.

## 8.2 Future works

### 8.2.1 Automatic indexing

Future research and applications of automatic indexing are, for example, to distribute the documents

with the keywords, the categories or the contents of the document, to summarize the sentences, to supplement the explanation of the keyword, to translate the keywords automatically and so on.

### 8.2.2 Information retrieval with association

The followings are items that need to be researched in order to achieve a higher Dr and Dp in information retrieval.

(i) Construction of a dynamic thesaurus

(a) Node weight calculating algorithm

The implemented system uses only the ranking of keywords, 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 each keyword.

(b) 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 subject-object relation. Furthermore, the link gener-

ation algorithm should generate various types of links, such as a location-link and an agent-link. Finally, all links should have a weight affixed to them.

(ii) 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, the degree of area activation, and so on.

References

[Abe 1985]
Abe,T., "The construction of ACE-CHUNICHI, a database of Chunichi newspaper articles," Information Management, Vol.28, No.2, pp.116-127, May, 1985(in Japanese).

[Akiyama 1988]
Akiyama,K., "The issue of intelligent information retrieval to textbase," Japanese Information Processing Society, SIG-DBS-64-3, 1988(in Japanese).

[Futamura et al. 1987]
Futamura,S. and Matsuo,F., "Automatic indexing by stop word removal on scientific and technical documents written in English," Transactions of Information Processing Society of Japan, Vol.28, No.7, pp.737-747, 1987(in Japanese).

[Hamill et al. 1980]
Hamill,K.A. and Zamora,A., "The use of titles for automatic document classification," Journal of the American Society for Information Science, Vol.8, No.3, pp.1-10, Nov., 1980.

[HAPPINESS]

References

HAPPINESS general information, HEIWA JOHOU CENTER CO., Ltd(in Japanese).

[Hayakawa et al. 1992]
Hayakawa.T. and Chida,T., "Automatic indexing system in the Kahoku Simpo Press - used JAIRS," The Journal of Information Science and Technology Association, Vol.42, No.11, pp.1033-1040, 1992(in Japanese).

[Hiroki 1981]
Hiroki,M., News Thesaurus, Kinokuniya Book Store, Tokyo, 1981(in Japanese).

[Ishikawa 1991]
Ishikawa,T., "A machine-aided subject indexing system based on sentence analysis," Transactions of Information Processing Society of Japan, Vol.32, No.2, pp.220-228, 1991(in Japanese).

[Ishikawa 1992]
Ishikawa,T., "Automatic indexing system on Japanese documents. The state-of-the-art report for the automatic indexing system to Japanese text data," Journal of Information Science and Technology Association, Vol.42, No.11, pp.994-1002, 1992(in Japanese).

[Ito 1991]

Ito,H., "Design and software structure of the knowledge-based information retrieval system NIRS," Trans. of IPSJ, Vol.32, No.9, pp.1102-1112, Sep., 1991(in Japanese).

[Jennings et al. 1992]
Jennings,A. and Higuchi,H., "A personal news service based on a user model neural network," Trans. of IEICE, Inf. & Syst., Vol.E75-D, No.2, pp.198-209, Mar., 1992.

[Jung et al. 1990]
Jung,G.S. and Raghavan,V.V., "Connectionist learning in constructing thesaurus-like knowledge structure," AAAI Symposium Text-Based Intelligent Systems Working Notes, pp.123-127, Mar., 1990.

[Kanou et al. 1991]
Kanou,Y. and Kishino,F., "An intelligent interface with a user model for document retrieval," Trans. IEICE, Vol.J74-D-1, No.8, pp.567-576, 1991(in Japanese).

[Kimoto 1987a]
Kimoto,H., "An automatic indexing system using linguistic processing," National Convention Record of IEICE, No.1450, p.6-128, Mar., 1987(in Japanese).

References

[Kimoto 1987b]
Kimoto,H., "An automatic indexing method using linguistic processing," Proc. of the First Annual Conference of JSAI, No.7-7, pp.389-392, 1987(in Japanese).

[Kimoto 1987c]
Kimoto,H., "Case vs. location characteristics of keywords in texts," Proc. of IPS Japan, No.5s-7, Sep., 1987(in Japanese).

[Kimoto 1987d]
Kimoto,H., "Automatic indexing and evaluation of keywords," SIG. Notes of Information Processing Society of Japan, NL., Vol.87, No.84, pp.1-8, 1987(in Japanese).

[Kimoto et al. 1989]
Kimoto,H., Nagata,M. and Kawai,A., "Automatic indexing system for Japanese texts," REVIEW of the Electrical Communications Laboratories, Vol.37, No.1, pp.51-56, 1989.

[Kimoto et al. 1990]
Kimoto,H. and Iwadera,T., "Construction of a dynamic thesaurus and its use for associated information retrieval," Proceedings of 13th International Conference on Research and Development in Information Retrieval, Brussels, Belgium, Presses Universitaires de Bruxelles,

pp227-240, Sep., 1990.

[Kimoto 1991a]
Kimoto,H., "Natural language processing and its application to Japanese database - the INDEXER system," Proceedings of the 3rd International Conference on Japanese Information in Science, Technology and Commerce, Vandoeuvre-les-Nancy, France, INIST, pp447-460, May, 1991.

[Kimoto 1991b]
Kimoto,H., "Automatic indexing and evaluation of keywords for Japanese newspapers," Transactions of IEICE Japan D-1, Vol.J74-D-1, No.8, pp.556-566, 1991(in Japanese).

[Kinukawa et al. 1982]
Kinukawa,H., Tanaka,K. and Ikegami,N., "Automatic indexing methods for information retrieval system of Japanese text," The HITACHI HYORON, Vol.64, No.5, pp.75-78, 1982(in Japanese).

[Luhn 1957]
Luhn,H.P., "A statistical approach to mechanized encoding and searching of literary information," IBM Journal of Research and Development, Vol.1, No.4, pp.309-317, 1957.

References

[Miyamoto et al. 1983]
Miyamoto,S., Miyake,T. and Nakayama,K., "Generation of a pseudothesaurus for information retrieval based on co-occurrences and fuzzy set operations", IEEE Trans. on Sys., Man, and Cybern., Vol.SMC-13, No.1, pp.62-70, Feb., 1983.

[Miyamoto et al. 1986]
Miyamoto,S. and Nakayama,K., "Fuzzy information retrieval based on a fuzzy pseudothesaurus," IEEE Trans. on Sys., Man, and Cybern., Vol.SMC-16, No.2, pp.278-282, Apr., 1986.

[Miyazaki 1984]
Miyazaki,M., "Automatic segmentation method for compound words using semantic dependent relationships between words," Transactions of Information Processing Society of Japan, Vol.25, No.6, pp.970-979, 1984(in Japanese).

[Miyazaki 1986]
Miyazaki,M. et al., "Linguistic processing method for a Japanese text to speech system," Transactions of Information Processing Society of Japan, Vol.27, No.11, pp.1053-1062, 1986(in Japanese).

[Morita,T. et al. 1989]

Reference

Morita,T. et al., "Fuzzy document retrieval system.(1) -- Experimental system and results," Proc. of IPSJ, No.2N-2, pp.1067-1068, Oct., 1989(in Japanese).

[Morita,Y. 1988]
Morita,Y., "An indexing scheme for terms using structural superimposed code words," ICOT Research Paper, No.383, pp.1-9, 1988.

[Nagao 1992]
Nagao,K., "Recent technical trends in natural language processing," Journal of Information Processing Society of Japan, Vol.33, No.7, pp.741-745, 1992(in Japanese).

[Nagao et al. 1976]
Nagao,M., Mizutani,M. and Ikeda,H., "An automatic method of extracting important words from Japanese scientific documents," Transactions of IPS Japan, Vol.17, No.2, pp.110-117, 1976(in Japanese).

[Nakazono et al. 1984]
Nakazono,K. and Shirai,S., "An automatic indexing system for Japanese text," Proc. of Symposium on Natural Language Processing Technologies, IPS Japan, pp.19-25, 1984(in Japanese).

[Ogawa et al. 1989]

References

Ogawa,Y. et al., "Fuzzy document retrieval system.(2)--A learning method of a keyword connection matrix," Proc. of IPSJ, No.2N-3, pp.1069-1070, Oct., 1989(in Japanese).

[Ohara 1991]
Ohara,H. et al., "Revision support techniques for Japanese text," NTT R&D, Vol.40, No.7, pp.905-914, 1991(in Japanese).

[Salton 1975]
Salton,G., A Theory of Indexing, 18 Regional Conference series in Applied Mathematics SIAM, J.W. Arrowsmith Ltd., Bristol 3, England, 1975.

[Salton et al. 1983]
Salton,G. and McGill,M.J., Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

[Salton 1987]
Salton,G., "Expert systems and information retrieval," ACM SIGIR Forum, Vol.21, No.3-4, 1987.

[Shibata et al. 1987]
Shibata,K., Miyanaga,Y. and Tochinai,K., "Automatic extraction of technical term information from scientific documents," 35th National Convention of Information

Processing Society of Japan, Vol.2, pp.1283-1284, 1987(in Japanese).

[Smeaton et al. 1988]
Smeaton,A.F. and Van Rijsbergen,C.J., "Experiments on incorporating syntactic processing of user queries into a document retrieval strategy," Proceedings of the 11th ACM Conference on Research and Development in Information Retrieval, Presses Universitaires de Grenoble, pp.31-51, 1988.

[Toriyama 1992]
Toriyama,T., "INDEXER utility on R&D information service system," The Journal of Information Science and Technology Association, Vol.42, No.11, pp.1017-1022, 1992(in Japanese).

[Yokoi 1993]
Yokoi,T., "A very large-scale knowledge base based on the amalgamation of knowledge processing and natural language processing -From electronic dictionaries to knowledge archives," Journal of Japanese Society for Artificial Intelligence, Vol.8, No.3, pp.286-296, 1993(in Japanese).

[Yoshimura et al. 1986]
Yoshimura,K., Hitaka,T. and Yoshida,S., "An automatic

References

extraction system of technical terms from Japanese scientific documents," Transactions of Information Processing Society of Japan, Vol.27, No.1, pp.33-40, 1986(in Japanese).