Automatic Index Generation For The Free-Text Based Database

A thesis
submitted to
The Department of Computer Science
The Chinese University of Hong Kong
in partial fulfillment of the requirements
for the degree of
Master of Philosophy

by

Leung Chi Hong

June 1992

Abstract

In this study, an automatic index generation method is proposed for the Chinese Medicinal Material Research Center in the Chinese University of Hong Kong. In this center, there is a free text-based database containing more than 10,000 documents about the Chinese Medicine. These documents are not yet indexed. The aim of this research is to develop the automatic indexing procedures to solve the problem of indexing documents in this center.

In this study, a statistical automatic indexing method is developed. The main reason of adopting the statistical approach is to by-pass the problems involved in the handling linguistic features of natural language. In this statistical approach, word occurrence frequencies and statistical correlations between indexes and words will be used to determine the index assignment for documents.

Many new ideas have been inspired in the study. First, there is a new method introduced to solve the problem of distinguishing correct index-word associations from incorrect ones. For words that are truly related to a certain index, these words are with lower chance to change their ranks which are determined by comparing the statistical correlations of words that are associated with the index. (Larger value of statistical correlation means higher rank.) When the number of indexed documents used to calculate these index-

word associations is increasing, there will be changes of word ranks. These are caused by the changes of statistical correlations which are approaching to the correct values when more and more documents are used to calculate these associations. But rank changes of correct words are relatively small compared with those of incorrect ones since increasing documents only establishes correct statistical correlations. This feature can be used to determine which index-word associations are correct. This method is better than the traditional method using only the statistical correlations between indexes and words (a large correlation value means correctness of an association). Because the statistical correlation may not always reflect the correctness of an association due to statistical errors and various relations between words and indexes. Second, the concept of word diversity is introduced in this paper. When categories of words found in a document are restricted and similar, the word diversity is low. Conversely, various and different word categories lead to high word diversity. The word diversity is an important factor affecting the performance of automatic indexing although it is seldom mentioned in the past researches. When the word diversities of documents are low, the performance of automatic indexing can be improved. Third, in this study, it is found that in a non-indexed document, the proportion of words proposing an certain index is correlated with the correctness of the index. A method using this feature is introduced to predict the correctness of proposed indexes automatically. Finally, the use of semantic representation for natural language terms has been attempted in this study and it is found that it can solve the problems in managing a large amount of natural language terms and representing synonyms and hierarchial-related

terms in the statistical approach of automatic indexing although this method is seldom used in the past researches for the statistical approach.

Simulations using imaginary data and case studies using real data have been performed to demonstrate that the procedures proposed in this paper can work practically to assist in automatic indexing of free-text documents.

Acknowledgements

I am deeply indebted to Dr. W. K. Kan, my supervisor of this research, for his continuing advice and encouragement through the years. He has been a constant source of knowledge and inspiration to me. Also, I wish to express my sincere thanks to Prof. T. C. Chen and Dr. Y. S. Moon for their valuable suggestions and corrections that contributed to the completion of this paper. Moreover, I would like to acknowledge the Department of Computer Science of the Chinese University of Hong Kong for providing assistance in this work. Finally, I am grateful to the staff of the Chinese Medicinal Material Research Center, especially Prof. H. M. Chang, Dr. P. H. But and Dr. C. M. Lee, for their collaboration and enthusiasm in this research.

Table of contents

Chapter one

Introduction

In the Chinese Medicinal Material Research Center (CMMRC) in the Chinese University of Hong Kong, there is a free text-based database containing more than 10,000 medical documents written in English text [2]. These documents are not yet indexed and searching information in this database is completely dependent on free text matching of query terms with text stored in the database. This searching method is so difficult that the searcher may need to attempt each possible clue word string to retrieve the information he wants.

The problem of this kind of searching is that there is no standard entry points to get access to the information stored in the database. One obvious solution to this problem is to classify documents in the database according to their subject contents. In other words, these documents should be indexed. Index terms will be assigned to each document to describe and summarize the document content. After the documents are indexed, the searching can be performed efficiently by using these index terms as query descriptions.

Now, the problems are (1) how a set of index terms for medical documents can be developed and (2) how these documents can be indexed by these index terms.

In the National Library of Medicine (NLM), a set of well-developed index terms for western medicine has been being used and modified for more than 100 years [9,10]. They are Medical Subject Headings (MeSH) which are used in the western countries as standard index terms for medical science. In fact, in the CMMRC database, the documents are talking about Chinese medicine treated by the western medical approach. Therefore, it is feasible that the MeSH index terms can be adopted in the CMMRC database.

But the second problem is how such a large amount of documents stored in the CMMRC database can be indexed. Typically, indexing is a task performed by a human indexer with certain knowledge and experience on the field where he works. It seems that human indexing may not be a practical solution to the problem of indexing these documents in this center. Automatic indexing is a feasible choice. It means the indexing task is assisted by the use of computer. The aim of this research is to develop an automatic indexing method to suit the circumstance of this center.

Automatic indexing for information expressed in the form of natural language is not a simple task. In the past, researchers attempted to tackle this

problem with different approaches. But there is still a room for improvement. An automatic indexing method based on statistical approach is developed in this research. Some new techniques have been attempted in the automatic indexing. These techniques include determining the correct associations between indexes and words, using semantic representation in the statistical approach and predicting the correctness of proposed indexes. The concept of word diversity will be introduced in this paper. This is a factor that is able to affect the indexing performance but is seldom noticed in the past researches. They will be described explicitly in this paper. Below are some brief descriptions of other chapters of this paper.

In the chapter two, the background knowledge about indexing will be covered. Two typical automatic indexing approaches using linguistic knowledge of natural language will be mentioned. Examples will be described in order to illustrate the concepts and techniques used in these approaches clearly. Comments on approaches using linguistic knowledge will be discussed.

In the chapter three, results of past researches on statistical approach of automatic indexing will be covered. The rationale of automatic indexing studied in this research will be described. The procedures of the automatic indexing and the solutions to the problems encountered in these procedures will be mentioned explicitly. Factors affecting the performance of automatic indexing will be studied.

In the chapter four, the results of simulations using imaginary data will be described. The aim of these simulations is to verify the automatic indexing method proposed in this paper. Factors affecting the performance will be considered in these simulations.

In the chapter five, real documents selected from the Chinese Medicinal Material Research Centre will be used to perform the procedures of automatic indexing. The aim is to illustrate the feasibility of using automatic indexing in the real world. Results and factors affecting the automatic indexing performance will be described.

In the chapter six, the findings and study results of this research will be concluded.

Chapter two

Background knowledge and linguistic approaches of automatic indexing

2.1 Definition of index and indexing

Indexes are a group of terms used to represent some special features of documents such as author and subject content. In other words, indexes are used to indicate the document content. Indexing is a process to assign suitable indexes for a document in order to describe the information carried by it.

The relation between indexing and searching is very close. The main aim of using indexes to describe a document is for searching. If the index assignment is not proper that indexes cannot reflect the document content and cannot be used for searching, indexes will become worthless. In other words, if in the searching process, there is no index to represent the document content, one needs to go through each document by examining its content.

Cleveland [3] listed different types of indexes. They are author indexes, subject indexes, classified indexes, coordinate indexes, permuted title indexes, faceted indexes, chain indexes, string indexes and citation indexes. Subject

index which reflects the subject content of the documents is one involved in this study and many past researches.

Rowley [17] divided indexing into controlled indexing and natural-language indexing based on the degree of control for using index terms. In the controlled indexing, only a set of predefined indexes can be used while in the natural-language indexing, any term of natural language can be used freely.

Indexing is a process requiring experience and knowledge. Cleveland [3] said that good indexing was not a causal clerical job, but the result of a professional activity carried out by people with proper training and experience. There are procedures and techniques, worked out over the years, that can be learned and followed.

2.2 Indexing methods and problems

Do the various methods used by librarians, documentalists and information scientists to organize knowledge and information keep pace with the growth of knowledge and our changing constructs of it ? Vickery [21] arranged the various methods of classification and indexing in an order of increasing degree of control. The list of these methods is shown below (arranged by increasing degree of control).

1. Words chosen from title or text, with common words omitted.
2. Words chosen from text, with omission of common words and consideration of variants.
3. Words chosen from text, with omission of common words, consideration of variants, and generic relationships.
4. Words chosen from text, with consideration of syntactical relationships between indexing terms.
5. Any of the preceding methods, with addition of terms not used in text.
6. Assignment of index entries from a fixed authority list or classification schemes.
7. Assignment of index entries from authority lists or classification schemes representative of several viewpoints and aspects of subject.

Similarly, Steinacker [19] classified indexing problem into several levels. He considered that the intellectual task of indexing has three problem levels as follows.

1. Selecting significant words or terms (phrases) from the text which are equivalent to thesaurus descriptors (consecutive or sequential indexing).
2. Referring very specific terms in order to reduce the variety of terms (hierarchic or generic indexing).
3. Choosing descriptors which neither occur in the text nor are indicated by more specific terms, but which are only implied (symbolic indexing).

2.3 Automatic indexing and human indexing

Referring to above problems, human indexing solves the problems of indexing on these three levels simultaneously without always clearly distinguishing between them. However, the difficulties encountered under operational aspects are (1) the natural inconsistency of human work which leads to some arbitrariness in assigning descriptors to documents and (2) the high cost and long time required, and the difficulty in finding qualified staff, for this kind of routine work.

Automatic indexing is defined to be a process in which indexes will be assigned to documents automatically with the aid of computers. The reason for

using computers is that documents can be processed with higher speed, higher consistency and lower cost. The automatic indexing can settle down problems of index assignment inconsistency, high cost and long time for human indexing. However, the main problem to be solved in automatic indexing lies in creating algorithms capable of identifying those elements of the text that can regarded as representatives of its contents. In most cases, indexing is a relatively easy decision for a human expert to make. The question being raised is whether a computer can be programmed to determine the subject content of a document and indexes which should be assigned to this document.

2.4 Different approaches of automatic indexing

There are three different approaches used in the automatic indexing of natural language. They are semantic, syntactic and statistical approaches. In fact, the first two approaches are ones related to linguistic knowledge of natural language. In this chapter, these two approaches will be introduced briefly while the statistical one will be mentioned in the next chapter.

Following are some typical examples using semantic and syntactic approaches. For semantic approaches, Vleduts-Stokolov [22,23] developed a formalized language which was used to match with natural language in the text of a document while Humphrey and Miller [7] used the frame-based knowledge representation language to assist in indexing process. Maeda [12] and Trubkin [20] used a dictionary containing some lexical knowledge and concepts to support the indexing process. Their common principle is that the text of a document will be interpreted in order to extract the semantic meanings carried by the linguistic entities such as words, phrases and sentences. Then, after analysis of these semantic meanings, the indexes reflecting these meanings will be assigned.

On the other hand, for syntactic approaches, Dillon and Gray [4] developed syntactical rules which were used to analyze the syntactical structure of the document text to extract suitable terms as indexes. Janas [8] used the

knowledge of linguistic regularities to recognize important phrases from the text while Sager [18] proposed a sublanguage grammer to extract information contained in the text for indexing. Their main principle is to locate content-bearing items (especially nouns and verbs) in sentences and then they will be used as indexes or as clues for choosing suitable indexes.

In order to illustrate the procedures and concepts used in these two approaches, a typical example for each approach will be explained concisely in the following paragraphs.

2.5 Example of semantic approach

In this example, the approach used by Vleduts-Stokolov [22,23] will be illustrated. Vleduts-Stokolov described a natural language processing system designed as an automatic aid to subject indexing in BIOSIS. The procedure that the system should model is a deep indexing with a controlled vocabulary of biological concepts -- Concepts Headings (CHs). On the average, ten CHs are assigned to each article by BIOSIS indexers.

The automatic procedure consists of two stages: (1) translation of natural-language biological titles into title-semantic representations which are in the constructed formalized language of Concept Primitives, and (2) translation of the latter representations into the language of CHs.

The first stage is performed by matching the titles against the system's Semantic Vocabulary (SV). The SV currently contains approximately 15,000 biological natural language terms and their translations in the language of Concept Primitives. Following are examples of simple SV structure.

natural language term: Fatty Tissue
Concept Primitives: BONES, JOINTS & ADIPOSE TISSUE; LIPIDS

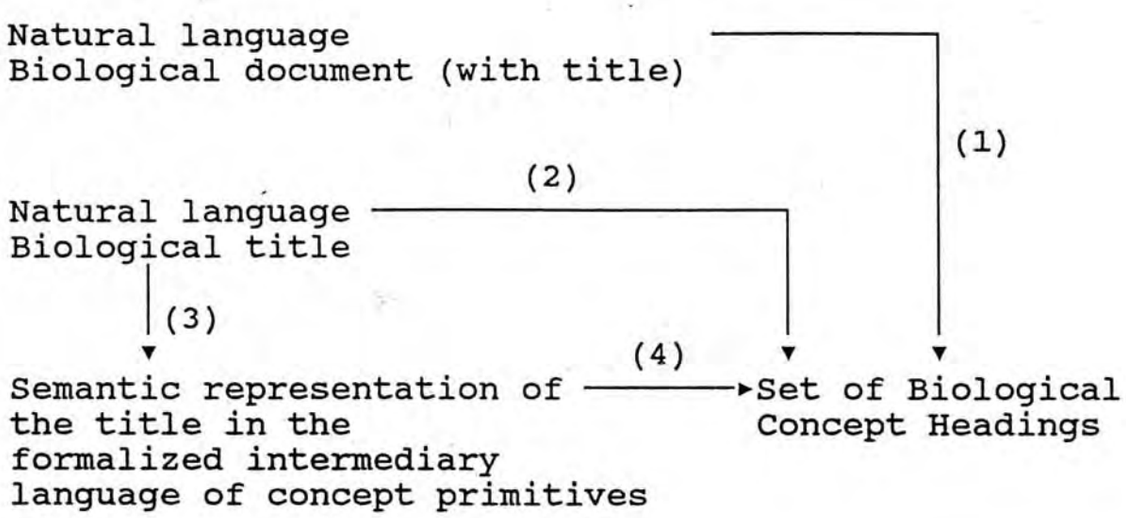
natural language term: Chicken(s)
Concept Primitives: AVES; PRODUCTION ANIMALS; POULTRY; TERRESTRIAL; DOMESTIC; LABORATORY

The second stage of the automatic procedure is performed by matching the title representations against the CH definitions, formulated as Boolean search strategies in the language of Concept Primitives. Following is an example of a CH definition.

CH: Blood Cell Studies
Definition: (CYTOLOGY) and (not THYMUS)

This semantic approach can be summarized in the table 1.

Table 1 Summary of semantic approach of automatic indexing used by Vleduts-Stokolov [21,22].



The arrow (1) corresponds to the real life indexing procedure. The arrow (2) corresponds to the virtual procedure, which is part of the procedure (1) modeled in the automatic procedure. The arrows (3) and (4) represent the two stages of automatic procedure.

2.6 Example of syntactic approach

In this example, the approach used by Dillon and Gray [4] will be illustrated. Their approach is based on the idea that content bearing words or phrases belong to certain syntactic categories or combinations of categories. After assigning the words in the text to categories, it selects concepts based on predefined patterns of categories. It then reduces variations of these concepts to an authoritative form for grouping. In practice, indexing consists of two major operations. The first is concept selection and the second is concept grouping.

The concept selection consists of three steps. The first step is assignment of words to syntactic categories. An exception dictionary of words and a suffix dictionary of word endings are used to assign mnemonic tags representing syntactic categories to every word, number, and punctuation character found in text. Since individual words within the English language may belong to more than one category, more than one tag may be assigned. Any word not tagged by the dictionaries is assigned a default tag of adjective-noun-verb. Following are some examples of tags.

Syntactic category	Examples
adverb or preposition	by, around
general noun	analysis
adjective	administrative
modal auxiliary	can, may

The second step is disambiguation of multiply tagged words. Choosing between multiple syntactic categories (disambiguation) is accomplished by examining the tags of words before and after the ambiguous (multitagged) word. For example, the word "automated" may be either a past tense verb or past participle. In the phrase "by automated methods", one rule for disambiguation recognizes that a past tense verb cannot follow "by", a word which functions either as a preposition or adverb, and the past tense tag is removed.

The third step is to select concepts. The text, represented by tags, is matched against a dictionary of acceptable concept forms. In the case of "by automated methods", the form identifies "automated methods" as a concept based on the tags "past participle" followed by "plural noun".

The second operation is concept grouping which is made up of two steps. The first step is formation of canonical forms. Each concept is first standardized by purging it of unwanted words, either general nouns, or words such as "by", "in", "of", "for" or "to". For example, "of" is purged from the phrase "review of books". Words for purging are identified by membership in syntactic categories. The remaining words of a concept (in stem form) are then sorted. The intent is to merge concepts that differ in minor ways to the same (canonical) form.

The second step is to group concepts. Quasisynonymous groups of concepts are formed by treating as equivalent all canonical forms that overlap in at least one stem.

After all these processes, content bearing words are selected and grouped according to their meanings. These words can be used as indexes or as clues for selecting indexes.

The methods used in this syntactic approach will be illustrated with an example in the table 2.

Table 2 An example illustrating the syntactic approach of automatic indexing used by Dillon and Gray [4]

A: CONCEPT SELECTION

A sample of text: I would like all information on library catalogs produced by automated methods ...

Tagging and disambiguation (steps 1-2)

Text	Tag	Dictionary	Disambiguated
I	PPS	Exception	
would	MD	Exception	
like	VB-SC-JJ	Exception	VB
all	PQL-QL	Exception	
information	GN	Exception	
on	APP	Exception	
library	NN	Exception	
catalogs	NNS-VBZ	Suffix	
produced	VBD-VBN	Suffix	
by	AP	Exception	
automated	VBD-VBN	Suffix	VBN
methods	NNS	Exception	

Concept selection (step 3)

Concept	Form
library catalogs	NN NNS-VBZ
automated methods	VBN NNS

B: CONCEPT GROUPING

Results of stemming, internal sorting of stems within the concepts and grouping multiword forms with single word groups.

Concept	Canonical Form	Groups
library catalogs	catalog librar	catalog librar catalog librar
automated methods	autom method	autom method autom method

2.7 Comments on semantic and syntactic approaches

Both the semantic and syntactic approaches imitate human using linguistic knowledge to understand the meaning of linguistic entities. Human also depends on this knowledge to understand the document content.

The main drawback of these approaches is that this knowledge used by the computers must be predefined first. One needs to define and store linguistic information about the natural language terms. In the semantic approach, semantic representations of each term have to been defined while in the syntactic approach, the syntactic roles possibly played by each term should be specified. Of course, to predefine such linguistic knowledge requires expertise and rather long time.

On the other hand, the success of these approaches relies on the correctness and completeness of the knowledge incorporated in the system. For instant, Vleduts-Stokolov has defined semantic representations for many natural language terms. If there is a new term not yet defined by semantic representations, this new term cannot be utilized in automatic indexing.

Chapter three

Rationale and methodology of automatic index generation

3.1 Problems caused by natural language

As mentioned in the previous chapter, the aim of automatic indexing is to determine indexes which are suitable to describe the content of a certain document with the aid of computers. In other words, the automatic indexing involves in the mechanical process of deciding what a certain document is talking about.

However, it is not easy to have a computer program which is able to understand the natural language as well as human does. As mentioned earlier, the most difficult problem in the automatic indexing is to deal with the linguistic features of natural language. It is because the computer program is not only required to identify the linguistic entities such as words, phrases and sentences, but is also required to interpret the meanings carried by these linguistic entities. The problem is complicated by the fact that there is no definite rule governing how these linguistic entities are combined to bear numerous types of meanings. Human can understand the meaning of the natural language, and then index the documents. Nevertheless, human also

depends on the experience and the knowledge of the natural language to grasp the meaning of natural language.

Therefore, as mentioned in the previous chapter, both the semantic and syntactic characteristics of the natural language have been utilized by many researchers in order to tackle this problem. In their methods, they developed knowledge components containing linguistic knowledge. The computer will make use of this predefined knowledge to interpret the meaning of the natural language and understand what the document is about, and finally suggest some indexes to it. But the success of their approaches is relied on the completeness and the correctness of the predefined knowledge incorporated in the automatic indexing procedures. Moreover, the development of such a knowledge component is rather time-consuming.

3.2 Usage of word frequencies

Is there any automatic indexing method which can by-pass the linguistic difficulties of understanding the semantic and syntactic structures which convey the meaning of the document ? In 1949, a book "Human Behaviour and the Principle of Least Effort" was published by George Zipf [24]. The main aim of his book was to support his thought of the principle of least effort. According to his principle, if there are many ways to achieve a goal, people will take the way requiring the least effort. He believed that this principle governed

many aspects of our activities including the use of language. In his book, he mentioned a behavioral factor related to word occurrence frequencies in English language texts. Zipf believed that after a length of time, people would be accustomed to the use of the least number of words to express the most meanings. His belief seems not so strange that everyday we tend to use comparatively few words out of the dictionary to express our thoughts and do the conversions.

Zipf calculated the occurrence frequencies of words in many texts. He finally made a conclusion that if words in a document are ranked according to their occurrence frequencies (the most high-frequency word has rank one, the second most high-frequency word has rank two and so on), the following relation is found.

Rank of word x occurrence frequency = constant

The above equation is the Zipf's first law. However, this law is only held when the word rank is high (ie. the occurrence frequency is large). Zipf also proposed another equation for words with low occurrence frequencies. This equation is as follows.

$$I_1/I_n = (4n^2-1)/3$$

where I_1 is the total number of words occurring one time and I_n is the total number of words occurring n times.

The Zipf's research on predictable behaviour of word frequencies initiated studies of statistical aspects of the natural language. Booth [1] proposed a different law for low-frequency words found in a document. This law is as follows:

$$I_1/I_n = n(n+1)/2$$

where I_1 is the total number of words occurring one time and I_n is the total number of words occurring n times.

Booth believed his law was more suitable to describe the characteristics of low-frequency words. Goffman [5] suggested that there should be a transitional region where the characteristics of high-frequency words following Zipf's first law will transform to those of low-frequency words following Booth's law. Goffman thought that the high-frequency words found in a document were functional words such as articles and prepositions which bear insignificant meanings while low-frequency words are ones reflecting the style and vocabulary diversity of the writer. Therefore, the medium-frequency words are ones carrying significant meanings and they can represent the main ideas of the document.

Pao [15] implemented Goffman's idea. According to Booth's law, to arrive at the transition point, words of low frequency will begin to take on the characteristics of words of high frequency. The number of words having n

frequency begins to approach unity (ie. $I_n \rightarrow 1$). Substituting one for I_n in the Booth's law, it becomes

$$I_1/1 = n(n+1)/2$$

Solving this equation,

$$n = \frac{-1+\sqrt{1+8I_1}}{2}$$

Therefore with the calculated value of n, one can easily identify the words around the transitional region. With this method Pao performed an experiment to locate the transitional region in some documents and he got a satisfactory result.

Luhn [11] thought that the frequency approach was sound. He believed when a writer wrote a document, he would select a comparatively small set of words used repeatedly to represent the major concepts of the document. Therefore, the words with certain high degree of occurrence frequencies can represent the main concepts. For example, in a document about the education, the occurrence frequencies of words such as "teacher", "school" and "examination" will be higher than those of words unrelated to this topic.

The common conclusion of these past researches is that there is a relation between occurrence frequencies of words and the document content.

Therefore, this feature can be used as a hint to suggest indexes for documents in the automatic indexing.

3.3 Brief description of rationale

The approach introduced in this paper is one which uses word frequencies rather than linguistic knowledge. The rationale and concepts used in this approach will be described first. Then, the details of procedures used in this method will be mentioned.

Indexes assigned to a certain document are used to reflect the concepts found in this document. But the occurrence frequencies of words (other than stopwords such as "and", "the") are also related to the document content. Therefore, there are relations between indexes and occurrence frequencies of words found in the document. If one can identify these relations between indexes and words, one can make use of these relations to index a document according to the word frequencies found in this document. Now, the problems are how these index-word relations can be identified and obtained, and how these relations can be used in the automatic indexing procedures.

The first problem is how the relations between words and indexes can be acquired. Obviously, these relations can be extracted from the documents that have been indexed already. The rationale of acquiring these index-word

relations is illustrated in the following example. Assume there are 100 indexed documents each of which contains a word X. Among these 100 documents, there are 50 documents and 25 documents indexed by an index A and an index B respectively. According to the statistics of these indexed documents, one can say that the word X has the chance of 50% to be with the index A, and only 25% with the index B. In other words, if there is a non-indexed document containing only a word X, the chances that it will be indexed by an index A and an index B will be 50% and 25% respectively.

In fact, in the above simplified example, one deals with a conditional probability that a document is indexed by a certain index, provided that a certain word is present in the document. If this probability is higher, the relation between the index and the word will be closer. Therefore, these relations are expressed in terms of statistical correlations between indexes and words found in the indexed documents. Based on this rationale, one can calculate the statistical correlation between each word and each index found in the indexed documents.

After the associations between indexes and words are acquired and expressed in the form of conditional probability, it will become ready to solve the second problem that how these index-word relations are used to index documents. After the relations between indexes and words have been developed, words found in a non-indexed document can be used to propose

indexes. Since each index may be related to two or more words, and likewise, each word may be related to two or more indexes, a list of candidate indexes with their corresponding probabilities is built. The probabilities of associations between indexes and words can be treated as proposing weights of indexes. The indexing process is achieved by adding each index's proposing weights suggested by all words found in the document, and then selecting some indexes with largest proposing weights.

The rationale of this approach based on word frequencies and index-word relations has been briefly described. The detail aspects of this automatic indexing method and solutions to the technical problems encountered in these procedures will be described in the following paragraphs explicitly.

3.4 Automatic index generation

Automatic indexing method proposed in this paper are mainly made up of two phases. They are training phase and indexing phase. In the training phase, a number of indexed documents (training documents) are analyzed. Based on these training documents, associations (ie. statistical correlations) between indexes and words of documents are searched and extracted out. In the indexing phase, words of a document (not yet indexed) will be analyzed and then indexes will be assigned to it in accordance with the words found in this document and index-word associations calculated in the training phase. The following paragraphs will describe procedures used in these two phases.

3.4.1 Training phase

The training phase are mainly divided into four processes. The first process is to select a number of indexed documents to form the training document set which will be used to calculate the associations between words and indexes. The second process is to control and standardize the variants of words found in the training document set. The third process, the main step in the training phase, is the calculation of associations between words and indexes. The final process is to discard some false associations between inappropriate indexes and words.

3.4.1.1 Selection of training documents

In the training phase, a number of indexed documents (training documents) will be selected for calculating index-word associations. There are some criteria used to select training documents. First, the contents of training documents should be related to those of documents that may be indexed in the future. Second, on the average, the occurrence frequencies of words and indexes found in all training documents should not be too low to cause statistical errors. These two requirements may be fulfilled simultaneously by selecting a large number of training documents. It is because when more training documents are used, more topics will be covered and the occurrence frequencies of words and indexes will also be higher.

3.4.1.2 Control and standardization of variants of words

For every training document, stopwords will be first eliminated because they bear insignificant meanings in the texts. Words such as "the", "of" and "when" are members of the stopword list. The complete stopword list is shown in Appendix A. In fact, for each knowledge domain, there should be an additional stopword list which is domain specific. For example, if the training documents are about medical science, words such as "drug" and "disease" will be very common in these documents. They have little importance to reflect the document contents and should be included in the domain specific stopword list.

After elimination of stopwords, the standardization of remaining words is usually followed. The aim of this step is to control and standardize the variants of words. There can be several levels of standardization. The first level is word stemming. Words of common origin will be treated as one word. For example, words "calculated", "calculating", and "calculation" can be counted as one word "calculate". In the second level of standardization, synonyms will be controlled. For example, the words "rifles", "pistols" and "shotguns" have the similar meaning of 'gun'. Different word forms of synonyms will be counted as occurrence of one standard form. The third level of standardization will be more complicated that some syntactic rules will be involved. In this level, the meanings of linguistic entities will be interpreted in order to control meanings conveyed by different combinations of linguistic entities. For example, both the word phrases "tree of apple" and "apple tree" contain the same meaning. These two phrases should be treated as identical phrase.

However, the second and third levels of standardization involve intensely in semantic and syntactic interpretation of natural language. They are often ignored in the automatic indexing based on word frequencies. For example, Hamill and Zamora [6] only used the first level of standardization while Maron [13] did not use the word standardization at all.

3.4.1.3 Calculation of associations between words and indexes

The main process in the training phase is the calculation of associations between words and indexes. Assume there are n different indexes $\{i_1, i_2, i_3, \ldots i_n\}$ and m different words $\{w_1, w_2, w_3, \ldots w_m\}$ found in all training documents after the stopword elimination and word standardization. The co-occurrence frequency of a word w_i and an index i_j, f_{ij}, is defined to be the frequency of a word w_i occurred in training documents indexed by an index i_j. The total occurrence frequency of a word w_i in all training documents is the sum of f_{i1}, f_{i2}, f_{i3}, ... f_{in}.

Thus, the occurrence frequency of a word w_i, F_i, is as follows.

$$F_i = \sum_{j=1}^{n} f_{ij} \quad (1)$$

The conditional probability, $P(i_j/w_i)$, that a document contains a word w_i and an index i_j, providing that the word w_i is present in the document, is given by

$$P(i_j/w_i) = \frac{f_{ij}}{F_i} = \frac{f_{ij}}{\sum_{j=1}^{n} f_{ij}} \quad (2)$$

The value of $P(i_j/w_i)$ for a word w_i and an index i_j represents the strength of association between them. If the $P(i_j/w_i)$ value is larger, the word w_i will be more related with the index i_j. The conditional probability $P(i_j/w_i)$ between each word and each index found in all training documents will be calculated to form a matrix that holds relations between indexes and words.

Using conditional probability is common in the statistical approach. Hamill and Zamora [6] had used the same conditional probability to express the statistical correlations between words and document categories in their automatic classification method while Maron [13] had used another conditional probability $P(i_j/w_a, w_b, w_c, \ldots w_n)$ to represent the statistical correlations between a certain class of document and a group of words simultaneously occurring in a certain document.

Although both the probabilities $P(w_i/i_j)$ and $P(i_j/w_i)$ can be derived from each other through Bayes' rule, the use of $P(i_j/w_i)$ is suitable to the circumstance that during the indexing phase words in a document are analyzed to propose suitable indexes. Also, Hamill and Zamora [6] suggested the use of $P(i_j/w_i)$ because it is easy to alter the size of the dictionary that contains words allowed to be used to calculate index-word associations.

It is emphasized that after the stopword elimination, there is no further selection of words which are allowed to be used to calculate the associations

with indexes. This approach is different from that used by Maron [13]. He rejected the words whose occurrence frequencies in training documents are either very high or very low. He claimed that the high-frequency words were too "common" to be clues for the specification of subject content while the low-frequency words are inefficient to be clues due to their rarities. For example, he said that high-frequency words such as "computer", "system" and "data" are too common in the general field of computers while all those words that appeared fewer than three times in the training documents will not be used to calculate the associations with indexes. Nevertheless, there is a drawback in his approach. It is difficult to determine which word frequency is "very high", "high", "low" and "very low". Maron also did not have any explicit and objective criteria to determine.

After the calculation of index-word associations, Maron [13] deleted some associations. If a certain word does not have a peak value in association with any index, all associations of this words will be deleted. Similarly, Hamill and Zamora [6] only retained associations with $P(i_j/w_i)$ values greater than or equal to 0.75. They used the $P(i_j/w_i)$ values to determine the importance (or correctness) of associations but there are some drawbacks. High $P(i_j/w_i)$ may be due to statistical fault caused by low occurrence frequencies of words in all training documents. Conversely, low $P(i_j/w_i)$ may not be necessary to mean false association since some words are actually related to many different concepts with equal importance. Values of $P(i_j/w_i)$ are divided and shared by

many indexes. For example, the word "base" may be related to the indexes about baseball, mathematics, military and chemistry.

But in fact, there are some unexpected associations calculated and included in the training phase. These unexpected (or false) associations should be discarded in order to increase the efficiency of the training phase. The technique used to determine false associations will be discussed in the following paragraph.

3.4.1.4 Discarding false associations

After $P(i_j/w_i)$ is calculated for every combination of word w_i and index i_j, each index i_j will be associated with m candidate words (ie. w_1, w_2, w_3, ... w_m) by m different $P(i_j/w_i)$ values respectively. However, for each index, only some words (ie. subset of $\{w_1, w_2, w_3, \ldots w_m\}$) are truly associated with it. Other associations are false.

Following is an example to illustrate the reason why there will be some false associations included in the training phase. Assume there is a document containing only three words (w_1, w_2 and w_3) and three indexes (i_1, i_2 and i_3). The correct associations should be i_1-w_1, i_2-w_2 and i_3-w_3. However, according to the procedures in the training phase, in this document each of three words will be linked with each of three indexes to form nine different associations.

As mentioned earlier, using the value of $P(i_j/w_i)$ to distinguish correct associations from incorrect ones has some drawbacks. After Hamill and Zamora [6] had used this parameter to determine correctness of associations, they found that there was a problem. They discovered that some words with large $P(i_j/w_i)$ were not useful for indexing process while some discarded words with small $P(i_j/w_i)$ were valuable to be retained.

In fact, a method independent of the value of $P(i_j/w_i)$ is required to distinguish correct associations from incorrect ones. There is a technique introduced to solve the problem mentioned above. This technique utilizes a characteristic that when the training document number is becoming larger, accurate associations between a certain index and corresponding words are being established more solidly. This means that at first (ie. few training documents) the associations are not yet accurately established. But these associations will be converged to proper structures when more and more training documents are being used. In the table 3, there is an example to illustrate the convergence of associations between an index and words when the number of training documents is increasing.

In the table 3, words in the associations are sorted by $P(i_j/w_i)$ with descending order. From this example, it is found that as the number of training document is increasing, the associations between the index i_1 and words w_1, w_2, w_3 and w_4 are becoming stable and accurate. Ranks (ie. order of words sorted

Table 3 An example illustrating the convergence of correct associations between an index and words.

Correct associations for a certain index i_1 are assumed to be as follows.

index	associated words
i_1	w_1, w_2, w_3, w_4

Convergence of associations to proper structures

	associations between i_1 and words
Training doc. no.=100:	i_1 <- w_1, w_3, w_5, w_2, w_6, w_4
Training doc. no.=500:	i_1 <- w_1, w_2, w_3, w_7, w_4, w_8
Training doc. no.=1000:	i_1 <- w_1, w_2, w_3, w_4, w_7, w_8
Training doc. no.=2000:	i_1 <- w_1, w_2, w_3, w_4, w_5, w_9

by $P(i_j/w_i)$ value) of words truly associated with the index will also become stable. As illustrated in this example, rank of w_2 is not changed when the training document number is 500 or more. In other words, as the training document number is increasing, change of word rank will become smaller and smaller if this word is truly associated with a certain index. It is because the increase in the training document number will only cause the proper structure of an association to be more accurately established. If a word is not truly associated with a certain index, the change of word rank will be comparatively large when the training document number is increasing.

According to this rationale, one can calculate and record a rank change, ΔR_{ij}, of a word w_i in an association with an index i_j each time when training

documents are increased. The sum of rank change, $\Sigma\Delta R_{ij}$, is defined to be summation of all ΔR_{ij} values which are recorded each time when training documents are increased. Thus, the smaller the $\Sigma\Delta R_{ij}$ value, the higher the probability that word w_i is truly associated with an index i_j.

Since words have different occurrence frequencies, the $\Sigma\Delta R_{ij}$ value of each word should be normalized by its occurrence frequency in all training documents. If F_i is the occurrence frequency of a word w_i in all training documents, the value of $(\Sigma\Delta R_{ij})$ will be normalized to $(\Sigma\Delta R_{ij})/F_i$. This normalization process is necessary because the higher the occurrence frequency, the higher the chance that the rank will be easily altered.

This new technique is an improved method to judge the correctness of the index-word associations. The advantage of using the rank change to determine the correctness of an index-word association is that this parameter will not be affected by the value of $P(i_j/w_i)$. The change of rank (determined by the comparison between words' $P(i_j/w_i)$ values) of a wrong word will be large, no matter what the $P(i_j/w_i)$ value will be. This method can avoid subjective determination of which range of $P(i_j/w_i)$ values reflecting correct associations.

The way of using rank change to discard false associations is rather straight forward. First, one can divide the training documents into several

portions. Then, one needs to perform the training phase each time after one portion is appended. The values of ΔR_{ij} are calculated along with each training phase calculation. Finally, after all training documents are used, the values of $(\Sigma\Delta R_{ij})/F_i$ can be obtained and one can sort the candidate words by $P(i_j/w_i)$ values (calculated in the last training phase) with descending order for each index. The obvious change of $(\Sigma\Delta R_{ij})/F_i$ value between two successive sorted words is a marker for detecting the boundary that separates correct associated words and incorrect ones. One can cut off those words which are listed after this boundary. The remaining words will be considered to have true associations with the index.

3.4.2 Indexing phase

The second phase in the automatic index generation is the indexing phase in which a number of indexes will be assigned to a non-indexed document based on (1) the word frequencies found in this document and (2) associations between words and indexes, established in the training phase.

For each non-indexed document, stopwords will be first eliminated and the standardization of remained words will be performed. These steps are same as those in the training phase.

Assume after the stopword elimination and the word standardization, there are certain words in a non-indexed document. These words will be used to propose corresponding indexes. For example, after the training phase, there are associations of two indexes i_1 and i_2 as follows.

index	associated words
i_1	w_1, w_2, w_3
i_2	w_2, w_3, w_4

If there is a non-indexed document with words w_2 and w_4 only, indexes i_2 will be proposed by words w_2 and w_4 simultaneously and the proposing weight of this index will be the sum of $P(i_2/w_2)$ and $P(i_2/w_4)$. The index i_1 will also be proposed. But only w_2 proposes this index whose proposing weight will be the value of $P(i_1/w_2)$.

The calculation of proposing frequency, PF_j, of a certain index i_j in a non-indexed document will be defined as follows.

$$PF_j = \sum_{i=1}^{m} f_i \times P(i_j / w_i) \quad (3)$$

where m is the number of different words occurred in the training phase and f_i is the frequency of a word w_i in a certain non-indexed document.

According to the above equation, the proposing frequency of an index i_j for a non-indexed document will be dependent on (1) statistical correlations between words and the index i_j and (2) frequencies of words found in this document. The rationale for this equation is that if a certain index is frequently proposed by words of a non-indexed document, this index will has a significant probability to be a suitable index for this document. This probability is reflected by the PF_j value of the index. Thus, for each non-indexed document, after calculating PF_j values of all indexes, indexes with comparatively high PF_j values will be selected and assigned to the document.

For documents to be indexed in the indexing phase, the PF_j value of a certain index i_j will be different from one document to another. It is because words (and their frequencies) that appear in each document are different from each other. Thus, the feature of a document can affect the selection of indexes

assigned to it. This is important in the automatic indexing in which the document content can be analyzed to determine appropriate indexes.

3.4.3 Example of automatic indexing

In the following paragraphs, an example will be used to illustrate the procedures used in the training and indexing phases. Assume there are four documents each of which has only one sentence. Those words underlined are significant words used to calculate index-word associations. The indexes are written in capital letters. These documents are listed below.

```
Doc 1
Index: VITAMIN, CARBOHYDRATE
Text: Vitamin B and starch are rich in rice.

Doc 2
Index: CARBOHYDRATE, FAT
Text: Starch and fat are rich in peanut.

Doc 3
Index: FAT, PROTEIN
Text: Fat and protein are rich in meat.

Doc 4
Index: VITAMIN, PROTEIN
Text: Vitamin B and protein are rich in fish.
```

The statistical correlation between an index VITAMIN and a word "starch", P(VITAMIN/"starch"), is as follows.

P(VITAMIN/"starch")
= co-occurrence frequency of VITAMIN and "starch"
occurrence frequency of "starch" in all training documents
= 1/2

Similarly, other index-word associations are calculated. They are shown on the table below. In this example, no discarding of false association is performed.

	"starch"	"protein"	"fat"	"vitamin B"
CARBOHYDRATE	2/2	0/2	1/2	1/2
PROTEIN	0/2	2/2	1/2	1/2
FAT	1/2	1/2	2/2	0/2
VITAMIN	1/2	1/2	0/2	2/2

Now, assume there is a non-indexed document, Doc X, as follows. Those words underlined are significant words used to propose indexes.

Doc X
Text: Protein, fat and vitamin B are rich in eggs.

According to the procedures of indexing phase, the proposing frequency of an index VITAMIN for Doc X will be as follows.

Proposing frequency of an index VITAMIN
= P(VITAMIN/"protein") + P(VITAMIN/"fat") + P(VITAMIN/"vitamin B")
= 1/2 + 0/2 + 2/2
= 1.5

Similarly the proposing frequencies of other indexes are calculated. The proposing frequencies of all indexes for Doc X are shown as below.

INDEX	Proposing frequency
CARBOHYDRATE	1.0
PROTEIN	2.0
FAT	1.5
VITAMIN	1.5

From the result, it is shown that the indexes PROTEIN, FAT, VITAMIN have higher probabilities to be assigned to the Doc X since they have higher proposing frequencies than the index CARBOHYDRATE.

3.5 Related researches

In the following paragraphs, the approaches used by Maron [13] and Hamill and Zamora [6] will be described since their approaches have some relationship with the problem encountered in this study.

Hamill and Zamora have developed an automatic classification system for chemical documents. In their case, the documents are required to be classified into one of eighty sections. The major techniques used in their study are like those used in the method described in this paper. Their approach involves in the calculation of the correlations between words of titles and classification sections found in some classified documents. These correlations will be used to suggest sections for non-classified documents. But in their study, there is a technique worth being mentioned. They have developed a heuristic routine that looks for chemical nomenclature roots. Chemical nomenclature is constructed from relatively few word roots that occur in many different combinations. Since many millions of substance names can be created in this way, it is not possible to achieve adequate dictionary matching for chemical nomenclature except for common substances. This word-root analysis allows assignment of words containing the specific chemical roots to appropriate sections. For example, the word root "PYRAZ" implies a ring system with two nitrogen atoms. Such substances are often found in one section which contains heterocyclic compounds with more than one hetero atom. Their method for

handling these chemical names can be applied in many disciplines which involves in the management of these chemical names.

Maron has also developed a system for classification of scientific papers. His approach also used statistical correlations between words and classes of documents. But in his approach, he has emphasized the use of key words which bear significant meanings related to the knowledge domain of classification. He first selected some key words from the text of the typical documents. Only these key words will be used to calculate the associations between words and classes of documents. His method is able to shorten the time requiring to perform calculation of class-word associations and assign a class for a document. It is because in these processes, fewer trivial words will be involved in the calculation process.

3.6 Word diversity and its effect on automatic indexing

In the following paragraphs, the concept of word diversity will be illustrated with some examples. The word diversity is an important factor affecting the whole performance of automatic indexing.

First, assume there are three sentences (S1, S2 and S3) made up of some imaginary words (a, b, c and d) as follows.

```
S1: a a a a
S2: a a b b
S3: a b c d
```

Each of these three sentences consists of four words. In S1, the number of different words found is one. For S2 and S3, there are two and four different words respectively. Thus, words of S3 have the highest degree of variance while those of S1 have the highest degree of similarity. One can say that S1 has low word diversity and S3 has high word diversity. Now consider the following two sentences (N1 and N2) of natural language.

N1: An apple pie is made from apples collected from apple trees.

N2: A machine is made up of parts brought from the factory.

If one only selects and considers nouns of these two sentences, N1 contains "apple pie", "apples" and "apple trees" while N2 contains "machine", "parts" and "factory". Nouns of N1 are about apple or apple-related. But nouns of N2 have no apparent relation with each other and they are more distinct. Thus, N1 has lower word diversity while N2 has higher word diversity. Following are two sentences (N3 and N4) of natural language.

N3: He likes apples and oranges.

N4: He likes apples and computers.

N3 can be modified to "He likes some fruits" since apples and oranges belong to the group of fruits. For N4, it is difficult to find a group of classification for apples and computers simultaneously. Thus, relation between apples and oranges is closer than that between apples and computers so that N3 has lower word diversity than N4. But if one compares N1 and N3, N1 has lower word diversity since relations between "apple pie", "apples" and "apple trees" are more closer.

Now, it is clear that the word diversity of a sentence involves in the measurement of the similarity of words. If the categories of words are restricted and similar, the sentence has lower word diversity. On the other hand, the word diversity will be higher if categories of words are various and different.

Sometimes, the determination of word diversity of a sentence or similarity between two words is rather subjective. Perhaps, one may feel that some sentences are not possible to be compared. For example, compare N2 and N4. It is difficult to tell which one has higher word diversity. If one says apples can be eaten while computers, machine, parts and factory cannot be eaten, N4 will has higher word diversity. But if one says factory is a place where things are made while machine, parts, apples and computers are things used by human, N2 will has higher word diversity. The problem is that there are often more than one criteria for comparison. It is emphasized that the criteria used to compare two words can affect the determination of similarity of them. For example, salary, income and wage may be the synonyms in the general cases but they have different and distinct meanings in the system for taxation. Some words have many different meanings that are related to different knowledge domains. For example, the word "base" is related to baseball, military, chemistry and mathematics. If a sentence contains two words "base" and "sport", the word diversity of this sentence is dependent on which concept of "base" is used to compared. If use the concept about baseball for the "base", the word diversity will be lower. But if use the concept about mathematics for the "base", the word diversity will be higher. One more example, the words "apple" and "computer" can have high similarity if one considers the word "apple" is a brand name of a computer manufacturer, "Apple Computer". In this case, the "apple" becomes computer-related.

Up to now, the word diversity is only concerned at the sentence level. In fact, the concept of word diversity can be described at higher levels such as paragraph, chapter, document, and so on. But when the level concerned is getting higher, the word diversity will be generally increased. For example, P1 is a paragraph of four sentences as follows.

P1: He has poor health and often gets influenza.
His doctor advises him to eat more fruits.
Fruits contain vitamin C against influenza.
Now he often eats apples and oranges.

If one considers the nouns (those words underlined) found in P1, the word diversity of P1 is higher than that of each individual sentence of P1. For example, the last sentence of P1 only deals with two fruit names (apples and oranges). But the complete paragraph of P1 contains fruit name, disease name (influenza), biochemical name (vitamin C) and career name (doctor). Thus, this sentence has lower word diversity than that of P1.

Assume P1 is followed by another paragraph P2 as follows. The paragraphs of P1 and P2 are supposed to form a small document.

P2: After eating apples and oranges for a long time,
he decides to grow fruit trees in his garden.
But he does not know how to start his plan.
He goes to library to look for information.

The word diversity of the whole document (made up of P1 and P2) will be higher than that of each paragraph in this document. If one considers only the nouns found in P1, it is relatively obvious to suggest linkage between nouns found in P1. For example, "doctor" is a person who can cure "influenza". But when the whole document (P1 and P2) is considered, the linkage for nouns found in the whole document will become relatively difficult to be discovered. For example, "doctor" is not working in the "garden" or "library". Thus, words "doctor", "garden" and "library" cause the word diversity of the document to be higher.

From the above examples, it is found that the word diversity at document level is higher than paragraph level, which in turn, is higher than sentence level. It is because when the level increases from the sentence level to document level, the categories of words will become more and more various and distinct. Thus, the word diversity is increased.

But in the automatic indexing system, one often concerns at the document level. Often, the index terms are assigned for a complete document but not for a paragraph or a sentence. Thus, in this study, the word diversity is concerned at the document level. In other words, word diversities of documents will be compared and studied. For example, if in a document, there are many occurrences of words such as "school", "teacher", "student" and "examination", the word diversity will be low since these words are related to

a topic of education. For another example, if in a document, there are many occurrences of words such as "computer", "chicken", "school" and "mountain", the word diversity will be higher than that of a previous example. At least, it is difficult to find a topic that is related to these words simultaneously.

As mentioned before, the determination of word diversity is rather subjective. Even if the words are known to belong to a certain knowledge domain, it is still difficult to compare the word diversities of two sentences. For example, in a document about computer science, there are two sentences. First sentence contains "computer", "software", and "data" while second sentence contains "processing unit", "programmer' and "database". It is hard to tell which sentence has lower word diversity. On the other hand, up to now, for two documents to be compared, we only consider that one document has higher word diversity than another one. But the degree of difference is not determined. The first step to solve the problem in calculating the word diversity and comparing the word diversities of documents objectively is to quantify the expression of the word diversity of a document. For example, the word diversity of the document X is 0.5 while that of the document Y is 0.8 so that the document Y has higher word diversity. There are two approaches to solve the problem in determining the similarity of two words and then the word diversity of a document.

The first approach uses the feature that if two words are identical, these two words should be associated with common indexes after the training phase in which associations between indexes and words are calculated. If two words w_a and w_b are associated with a set of common indexes, these two words can be treated as synonyms. For example, in the indexing system of food science, after the training phase, the words "rice", "wheat" and "grape" are determined to be associated with some indexes as follows.

Words	Indexes
rice	STARCH, CEREAL, MAKING-WINE
wheat	STARCH, CEREAL, MAKING-BREAD
grape	FRUCTOSE, FRUIT, MAKING-WINE

There two indexes commonly associated by "rice" and "wheat". But one index is commonly associated by "rice" and "grape" and no index is commonly associated by "wheat" and "grape". Thus, the order of similarity is as follows: rice-wheat > rice-grape > wheat-grape. However, each index-word association has a certain correlation value which should be considered in the calculation of similarity of two words. Assume there are n indexes covered in the training phase. After the training phase, the number of indexes associated with each word is n. (If there is no correlation between an index and a word, this can be represented by $P(i_j/w_i)$ equal to zero.) Each word with n $P(i_j/w_i)$ values can be treated as a vector in n-spaces. The similarity of two words can be represented

by the cosine correlation of these two vectors. Assume two vectors in t-spaces are $X = (x_1, x_2, \ldots x_t)$ and $Y = (y_1, y_2, \ldots y_t)$. The cosine correlation is

$$r = \frac{X \cdot Y}{|X|\,|Y|} = \cos\theta$$

$$r = \frac{\sum_{i=1}^{t} x_i y_i}{\sqrt{\sum_{i=1}^{t} (x_i)^2 \cdot \sum_{i=1}^{t} (y_i)^2}}$$

where Θ is the angle between vectors X and Y.

Thus, the similarity, S_{ab}, of two words w_a and w_b is defined as follows.

$$S_{ab} = \frac{\sum_{j=1}^{n} P(i_j/w_a) \cdot P(i_j/w_b)}{\sqrt{\sum_{j=1}^{n} P(i_j/w_a)^2 \cdot \sum_{j=1}^{n} P(i_j/w_b)^2}} \qquad (4)$$

The word diversity of a document can be defined in terms of the word similarity. Assume in a document, there are N words. For a word (say word X), it has (N-1) word similarity values with other (N-1) words. The average of these (N-1) word similarity values can represent the overall similarity between the word X and all other words of the document. For each word, this overall similarity is different. The average of these N overall similarities can be used to represent the word diversity of a document. According to this rationale, the average word similarity (AWS) is defined to reflect the word diversity of a

document. The AWS is defined as follows. The higher the AWS, the lower the word diversity of a document will be.

$$AWS = \frac{\sum_{i=1}^{N} \frac{\sum_{j=1, i \neq j}^{N} S_{ij}}{N-1}}{N} \quad (5)$$

But this approach relies on the calculation result in the training phase, which in turn, depends on the training documents selected and used in the this phase.

In the second approach of calculating similarity of two words and word diversity of a document, the significant meanings of words are predefined in advance in order to clarify the differences between the words deliberately. In this approach, each word of natural language is defined in terms of concept headings which are used to state the word's semantic meanings significant in a certain indexing system. For example, in an indexing system of animals, some words are predefined as follows.

Words	Concept headings
cattle	DOMESTIC, TERRESTRIAL, MILK PRODUCTION
cow	DOMESTIC, TERRESTRIAL, MILK PRODUCTION
hen	DOMESTIC, TERRESTRIAL, EGG PRODUCTION
duck	DOMESTIC, AQUATIC, EGG PRODUCTION

First, the concept headings of cattle and those of cow are exactly the same. Thus, these two words are treated as synonyms in this system. There are two concept headings commonly shared by cow and hen but only one is commonly shared by cow and duck. Therefore, the order of similarities of these word pairs is as follows: cow-cattle > cow-hen > cow-duck.

The mathematical way to represent the similarity, S^c_{ab}, of two words w_a and w_b, in terms of concept headings, is defined as follows.

$$S^c_{ab} = \frac{c_{ab}}{\sqrt{c_a \times c_b}} \qquad (6)$$

where c_{ab} is number of concept headings commonly shared by w_a and w_b while c_a and c_b are number of concept headings shared by w_a and w_b respectively.

In fact, the above equation is a simplified version of the cosine correlation. It is assumed that there is no weight assigned to concept headings. One only considers whether a certain word is described by a certain concept heading. In other words, weight of a certain word-concept-heading association is one or zero. Thus, one only needs to count how many concept headings are used to describe a word and how many concept headings are commonly shared by two words. Similarly, the AWS can be used to represent the word diversity of a document in this approach but S_{ab} is replaced by S^c_{ab} in the equation 5.

These two approaches of calculating word diversity have some common and different features. For common feature, both methods compare the characteristics of words. The first approach compares the degree of commonness of indexes associated by words while the second approach compares the degree of commonness of concept headings shared by words. For different features, in the first approach, word diversity depends on the calculation result of the training phase while in the second approach, word diversity depends on the expertise used to define concept headings for each word. The reliability of the first approach is related to the performance of training phase that can be implemented readily and inexpensively while that of the second approach is related to the expertise which requires long time and high cost to define the words in terms of concept headings.

Now, the effect of word diversity of the documents in the training phase and indexing phase will be mentioned. Generally speaking, the lower the word diversity, the higher the performances of these phases will be.

In the training phase, if the word diversities of the training documents are higher, there will be higher chance for false associations to occur. For example, assume there is a document mentioning something about mosquitos and this document is indexed by indexes MOSQUITOS and INSECT. If in this

document, there is a sentence "The mosquitos attacked with the ferocity of a tiger", some index-word associations made by this sentence are as follows.

```
Index-word associations
MOSQUITOS-"mosquitos"
INSECT-"mosquitos"
MOSQUITOS-"tiger"
INSECT-"tiger"
```

Obviously, the last two associations are falsely formed by this sentence. The word "tiger" in a document talking about mosquitos has increased the word diversity of this document. The metaphorical feature of natural language is a typical feature causing the high word diversity which leads to false associations. Sometimes the false associations caused by high word diversity of a document is inevitable. For example, assume there is an article talking about the Japanese food and this article is indexed by indexes JAPANESE CULTURE and JAPANESE FOOD. In this article, there is a sentence "Although some parts of culture of Japan originated in ancient China, SuShi is a typical kind of foods with Japanese style entirely". Assume country names are significant in the indexing system. According to this sentence, some index-word associations are formed as follows.

```
Index-word associations
JAPANESE CULTURE-"Japan"
JAPANESE FOOD-"Japan"
JAPANESE CULTURE-"Ancient China"
JAPANESE FOOD-"Ancient China"
```

Obviously, among these index-word associations, the last index-word association is falsely formed. The term "ancient China" occurring in an article talking about Japanese food has caused the word diversity to be increased.

From the above examples, it is found that the word diversity of each training document is an important factor to determine the number of false associations. If the word diversities of training documents are lower, the performance of this phase can get better.

Now the effect of word diversity on the indexing phase performance will be considered. Generally speaking, if the word diversity of a non-indexed document is lower, the indexes proposed for it will be more accurate.

For example, in a non-indexed document talking about vitamins, there is a sentence "Sunlight can stimulate the production of vitamin E but ultraviolet wave in the sunlight can cause harmful effect to human". The word "ultraviolet wave" is often found in the subject about physics but is relatively rare in the subject about life science. Therefore, the occurrence of this term can propose some indexes related to physics rather than life science. Thus, the physics-oriented term in a document about life science can increase the word diversity of the document and, thus, increases the chance to propose false indexes.

Sometimes the ambiguity feature of the natural language can increase the word diversity of a document inevitably. For example, in a chemical document talking about reaction between acid and base, there is a sentence "Acid can neutralize base". The word "base" can, in fact, propose some indexes related to other knowledge domains such as baseball, military and mathematics. In this example, these false indexes are proposed together with the expected indexes about chemistry. The multi-discipline-oriented terms can often increase the word diversity of a document.

From the above examples, it is found that when the word diversity is higher, there will be many words of different categories. They can propose different indexes related to their own categories separately. Thus, the chance of getting false indexes will be increased.

In conclusion, the word diversity of documents (training documents or non-indexed documents) is a critical factor affecting the performance of automatic indexing. The low word diversity can enhance the performances of both the training phase and indexing phase.

3.7 Factors affecting performance of automatic indexing

There are two factors which can affect the performance of the training phase. First is the size of the training document set and how many topics are covered by these training documents. If the number of training documents is large and more topics are mentioned in these documents, the associations between words and indexes will be more accurate. Because larger number of training documents can reduce the statistical errors caused by low occurrence frequencies of words and indexes. The second factor is the degree of diversity of words found in a document. When the categories of words found in a document are restricted and similar, the word diversity of the document is low. If the categories are various and different, the word diversity is high. Generally speaking, when the word diversities of training documents are low, the performance of the training phase is better.

There are two factors affecting the performance of the indexing phase. Of course, the first major factor is the accuracy of the associations between words and indexes, which are calculated in the training phase. The second factor is the word diversity of a non-indexed document. Lower word diversity of a non-indexed document can reduce the chance of proposing false associations.

3.8 Application of semantic representation

3.8.1 Problem of natural language

As mentioned earlier, one of the factors affecting the performance of the training phase is the size of the training document set. One of the aims of using a large training document set is to cover enough words and indexes to develop the correct associations between them. The number of indexes allowed to be used can be controlled willingly. In the NLM, the number of MeSH index terms used to index documents is around 16,000 [9,10]. One can easily check whether a certain index is already covered in the training phase. However, it is comparatively difficult to cover all nature language terms. If a non-indexed document contains some words not yet covered in the training phase, these words are unable to suggest indexes since they have no association with any index.

Even if one can cover all natural language terms, it will become difficult to manage such a large amount of terms. For example, one needs to use a large database to hold these natural language terms and their associations with indexes. Also, in the indexing phase, the time required to search such a large database will be rather long.

Another problem of the natural language is that some words have same or similar meanings but they are counted as different words. For example,

'rifles', 'pistols' and 'shotguns' have the similar meaning of 'gun'. Moreover, some words have shown hierarchial structures. For instance, 'taxi', 'train', 'bus' may be hierarchically under the concept 'transport vehicle'. If the synonyms and hierarchial structures of words can be represented and utilized in the automatic indexing procedures, the indexing performance can be increased in a certain extent.

3.8.2 Use of concept headings

In order to improve this circumstance, a method is introduced to attempt to solve the problems caused by the use of natural language. The natural language terms will be represented by a set of concept headings. The concept headings represent complex subjects and the meanings of the majority of them are combinations of several more "elementary" meanings. Following are some simple examples of concept headings suggested by Vleduts-Stokolov [22,23].

Natural language terms	Semantic representations in concept headings
apple(s), apple tree(s)	1. DICOTYLEDONS; 2. TEMPERATE ZONE FRUIT 3. TERRESTRIAL
apple juice	1. FOOD PRODUCT; 2. FOOD PROCESSING; 3. TEMPERATE ZONE FRUIT
apple moths	1. LEPIDOPTERA 2. PEST 3. TERRESTRIAL

With this approach a predefined set of concepts headings will be developed. Each important word of natural language will be represented by several concept headings simultaneously. In fact, using concept headings is a way to state clearly a term's semantic meanings which are essential to be identified in a certain indexing system. This means that concept headings of a certain word in two different indexing systems may not be identical. For example, in the system of food science, the concept headings assigned to the word "apple" may be significant different from those assigned to the word "orange" in order to clarify differences between them. But in the system of general science, the concept headings assigned to the word "apple" and "orange" may be same because treating them as identical thing "fruit" is already suitable for the indexing propose in this system.

Originally, the use of concept headings (or semantic representation) is very common in the automatic indexing with semantic approaches such as one used in Vleduts-Stokolov [22,23] but is seldom in statistical approach. In this paper, the concept headings will be attempted in order to investigate the feasibility of using them to suit the automatic indexing procedures mentioned in this paper.

The use of concept headings in the statistical approach is only with little change in the procedures. In the training phase, the words found in each training document will be converted into corresponding concept headings.

Then, the statistical correlations between indexes and concept headings will be calculated as those between indexes and words. In the indexing phase, the words found in each non-indexed document will be converted into corresponding concept headings first. Then, based on the presences and frequencies of these concept headings, the indexes will be proposed as those proposed by using words. On the whole, the main difference is that the automatic indexing procedures use concept headings instead of words to perform both the training phase and indexing phase. The procedures used for words also work for concept headings.

Assume in a document X, there is a sentence "The apple is used to make apple juice". The words underlined in this sentence will be converted into following concept headings (according to Vleduts-Stokolov's definition): DICOTYLEDONS, TEMPERATE ZONE FRUIT, TERRESTRIAL, FOOD PRODUCT, FOOD PROCESSING, TEMPERATE ZONE FRUIT. If the document X is a training document, these concept headings will be used to calculate the statistical correlations with the indexes assigned to the document X. If the document X is a non-indexed document, the frequencies of these concept headings will be used to propose indexes.

3.8.3 Example of using concept headings in automatic indexing

In this example, the use of concept headings in automatic indexing will be illustrated. Assume there are four training documents each of which has only one sentence. Those words underlined are significant words used in the training phase. The indexes are written in capital letters. These documents are listed below.

Doc 1
Index: VITAMIN, CARBOHYDRATE
Text: Vitamin B_1 and starch are rich in rice.

Doc 2
Index: CARBOHYDRATE, FAT
Text: Starch and linoleic acid are rich in peanut.

Doc 3
Index: FAT, PROTEIN
Text: Stearic acid and myosin are rich in meat.

Doc 4
Index: VITAMIN, PROTEIN
Text: Vitamin B_2 and collagen are rich in fish.

In this example, the words will be converted into corresponding concept headings. Then, these concept headings are used to calculate associations with indexes. The concept headings for significant words (those underlined in the text) are listed below.

Words	Concept headings
"vitamin B_1"	vitamin-B, anti-paralysis-chemical
"vitamin B_2"	vitamin-B, respiration-chemical
"starch"	polysaccharide, plant-energy-store
"linoletic acid"	fat, plant-fatty-acid
"stearic acid"	fat, adipose-tissue-constituent
"myosin"	protein, contraction-chemical
"elastin"	protein, structural-chemical

The statistical correlation between a concept heading plant-energy-store and an index FAT will be calculated as follows.

$$P(\text{FAT}/\text{plant-energy-store}) = \frac{\text{co-occurrence frequency of FAT and plant-energy-store}}{\text{occurrence frequency of plant-energy-store in all training documents}} = 1/2$$

Similarly, the statistical correlations between other indexes and concept headings are calculated. All these statistical correlations are shown on the following table.

	polysac-charide	plant-energy-store	fat	plant-fatty-acid	adipose-tissue-constituent	protein	contraction-chemical	structural-chemical	vitamin-B	anti-paralysis-chemical	respiration-chemical
CARBO-HYDRATE	2/2	2/2	1/2	1/1	0/1	0/2	0/1	0/1	1/2	1/1	0/1
PROTEIN	0/2	0/2	1/2	0/1	1/1	2/2	1/1	1/1	1/2	0/1	1/1
FAT	1/2	1/2	2/2	1/1	1/1	1/2	1/1	0/1	0/2	0/1	0/1
VITAMIN	1/2	1/2	0/2	0/1	0/1	1/2	0/1	1/1	2/2	1/1	1/1

Now, assume there is a non-indexed document, Doc X, as follows.

Doc X
Text: Globulin, cholesterol and vitamin B_5 are rich in eggs.

The words underlined are significant words used to propose indexes. These words will be converted into corresponding concept headings first. Then, based on these concept headings, indexes will be proposed according to the calculated index-concept-heading associations. The concept headings for words found in the Doc X are as follows.

Words	Concept headings
"vitamin B_5"	vitamin-B, anti-gut-disorder-chemical
"cholesterol"	fat, steroid-hormone-source
"globulin"	protein, antibody

The proposing frequency of an index PROTEIN for Doc X is as follows.

Proposing frequency of an index PROTEIN
= P(PROTEIN/vitamin-B) + P(PROTEIN/anti-gut-disorder-chemical) + P(PROTEIN/fat) + P(PROTEIN/steroid-hormone-source) + P(PROTEIN/protein) + P(PROTEIN/antibody)
= 1/2 + 0 + 1/2 + 0 + 2/2 + 0
= 2.0

Similarly, other indexes' proposing frequencies are calculated. All indexes' proposing frequencies for Doc X are shown as below.

INDEX	Proposing frequency
CARBOHYDRATE	1.0
PROTEIN	2.0
FAT	1.5
VITAMIN	1.5

3.8.4 Advantages of concept headings

The use of concept headings are with some advantages. First, the number of concepts headings allowed to use can be controlled readily and this number will be smaller than that of natural language terms. This can solve the problem of managing a large number of natural language terms.

Second, the synonyms of words and hierarchial relationships between words can be represented by the concept headings. If two words are synonyms of each other, their concept heading representations will be identical. If two words are hierarchically related with each other, concept headings representing the general word will also be used to represent the specific word.

Third, if a word in a non-indexed document is not yet covered in the training phase, this word can still be used to suggest indexes if synonyms or hierarchial-related words of this word have been covered already in the training phase.

3.8.5 Disadvantages of concept headings

There are some drawbacks of using concept headings. First, each important term of natural language should be defined in terms of concept headings in advance. This step is very time consuming. Also, the problems encountered in the process of defining these natural language terms will be similar to those encountered in the automatic indexing methods based on the sematic approach. This means that expertise and long development time are required.

Second, if one adopts concept headings in automatic indexing procedures, he will take a risk that the performance of the indexing phase may be deteriorated but not improved. The factor controls whether the performance is actually worsen is, again, the word diversity of a non-indexed document. Generally speaking, when the word diversity is low, the performance decline will not be apparent and will not affect the selection of correct indexes at all. Conversely, if the word diversity is high, the effect of performance decline will be significant. The reason of performance decline when the concept headings are used will be explained with some examples in following paragraphs.

Assume there are three imaginary documents. Each has three indexes and three words. Each word is represented by three concept headings. They are shown in the table 4.

Based on these three documents, association strengths (ie. $P(i_j/w_i)$) between indexes and words, and those between indexes and concept headings are calculated respectively. They are shown in the table 5.

Now, the documents listed in the table 4 will be treated as non-indexed documents, the words and concept headings are used to propose indexes based on the associations listed in the table 5. Two different sets of index proposing frequencies are given by using words and using concept headings respectively. The results are shown in the table 6.

Table 4 Imaginary documents each of which has three indexes and three words which are represented by three concept headings

Doc. 0	
i_0 i_1 i_2	<-- indexes
w_0 (c_0 c_1 c_2)	<-- word and concept headings
w_1 (c_1 c_2 c_3)	
w_2 (c_2 c_3 c_4)	
Doc. 1	
i_1 i_2 i_3	
w_1 (c_1 c_2 c_3)	
w_2 (c_2 c_3 c_4)	
w_3 (c_3 c_4 c_5)	
Doc. 2	
i_2 i_3 i_4	
w_2 (c_2 c_3 c_4)	
w_3 (c_3 c_4 c_5)	
w_4 (c_4 c_5 c_6)	

Table 5 Association strength tables calculated from documents listed in the table 4

Index-Word Association Strength Table

	w_0	w_1	w_2	w_3	w_4
i_0	1.00	0.50	0.33	0.00	0.00
i_1	1.00	1.00	0.67	0.50	0.00
i_2	1.00	1.00	1.00	1.00	1.00
i_3	0.00	0.50	0.67	1.00	1.00
i_4	0.00	0.00	0.33	0.50	1.00

Index-Concept Headings Association Strength Table

	c_0	c_1	c_2	c_3	c_4	c_5	c_6
i_0	1.00	0.67	0.50	0.29	0.17	0.00	0.00
i_1	1.00	1.00	0.83	0.71	0.50	0.33	0.00
i_2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
i_3	0.00	0.33	0.50	0.71	0.83	1.00	1.00
i_4	0.00	0.00	0.17	0.29	0.50	0.67	1.00

Table 6 Index proposing frequencies suggested by using words and using concept headings found in documents listed in the table 4

Doc. 0

indexes:	i_0	i_1	i_2	i_3	i_4
PFw:	0.20	0.30	0.33	0.13	0.04
PFc:	0.17	0.28	0.33	0.16	0.06

Correct indexes	(PFw-PFc)/PFw
i_0	0.150
i_1	0.067
i_2	0.000

Doc. 1

indexes:	i_0	i_1	i_2	i_3	i_4
PFw:	0.09	0.24	0.33	0.24	0.09
PFc:	0.11	0.23	0.33	0.23	0.11

Correct indexes	(PFw-PFc)/PFw
i_1	0.042
i_2	0.000
i_3	0.042

Doc. 2

indexes:	i_0	i_1	i_2	i_3	i_4
PFw:	0.04	0.13	0.33	0.30	0.20
PFc:	0.06	0.16	0.33	0.28	0.17

Correct indexes	(PFw-PFc)/PFw
i_2	0.000
i_3	0.067
i_4	0.150

PFw = Normalized proposing frequency of an index proposed by words

PFc = Normalized proposing frequency of an index proposed by concept headings

From this example, it is found that when concept headings are used to propose indexes, on the average the proposing frequencies of correct indexes will be reduced while those of incorrect indexes will be increased. The value of (PFw-PFc)/PFw can be used to measure the change of proposing frequencies

of correct indexes. If this value is large, the extent of the performance decline will be large.

When concept headings are used to associate with indexes, one "index to word" association will become several "index to concept heading" associations. After this transformation, it is not possible to identify which "index to word" associations contribute to a certain "index to concept heading" association. It is because one concept heading may be shared by two words or more.

Assume there is a non-indexed document, document D. The document D has some words which will be converted into some concept headings including a concept heading C. Also, assume the document D should not be indexed by an index X. However, in the training phase, some training documents (other than document D) indexed by the index X contain some words which will be converted into concept heading C. Therefore, an association between the concept heading C and the index X is constructed in the training phase. Nevertheless, in the indexing phase, the words (after converting into concept headings) of the document D will unintentionally propose the index X, a wrong index to this document.

For instance, in the Doc. 2 in the table 4, only word w_2 will suggest the wrong index i_0 if associations between words and indexes are used.

Nevertheless, if associations between concept headings and indexes are used, in the Doc. 2, word w_3 will be converted into concept heading c_3, c_4 and c_5, and word w_4 will be converted into concept heading c_4, c_5 and c_6. But both c_3 and c_4 have associations with the index i_0. Therefore, in Doc. 2, all words have suggested this wrong index i_0. But, in fact, the associations between c_3, c_4 and i_0 are constructed by other training documents containing the index i_0 and words which share the concept headings c_3 and c_4. For example, the Doc. 0 contains index i_0 and word w_2 which shares c_3 and c_4. Since the proposing frequencies of incorrect indexes are increased, the normalized proposing frequencies of correct indexes will be relatively reduced.

Following is another example showing the decline of the indexing phase performance. But in this example, the word diversity of the documents will be even higher. (The similarity of two words can be measured by the equation 6 and the word diversity of a document can be measured by the equation 5.) The documents are listed in the table 7.

Table 7 Imaginary documents with higher word diversity compared with documents listed in the table 4

Doc. 0
i_0 i_2 i_4
w_0 (c_0 c_1 c_2)
w_2 (c_2 c_3 c_4)
w_4 (c_4 c_5 c_6)

Doc. 1
i_1 i_3 i_5
w_1 (c_1 c_2 c_3)
w_3 (c_3 c_4 c_5)
w_5 (c_5 c_6 c_7)

Doc. 2
i_2 i_4 i_6
w_2 (c_2 c_3 c_4)
w_4 (c_4 c_5 c_6)
w_6 (c_6 c_7 c_8)

The association strengths calculated from these documents in the table 7 are listed in the table 8. The index proposing frequencies are listed in the table 9.

Table 8 Association strength tables calculated from documents listed in the table 7

Index-Word Association Strength Table

	w_0	w_1	w_2	w_3	w_4	w_5	w_6
i_0	1.00	0.00	0.50	0.00	0.50	0.00	0.00
i_1	0.00	1.00	0.00	1.00	0.00	1.00	0.00
i_2	1.00	0.00	1.00	0.00	1.00	0.00	1.00
i_3	0.00	1.00	0.00	1.00	0.00	1.00	0.00
i_4	1.00	0.00	1.00	0.00	1.00	0.00	1.00
i_5	0.00	1.00	0.00	1.00	0.00	1.00	0.00
i_6	0.00	0.00	0.50	0.00	0.50	0.00	1.00

Index-Concept Headings Association Strength Table

	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
i_0	1.00	0.50	0.50	0.25	0.40	0.25	0.25	0.00	0.00
i_1	0.00	0.50	0.25	0.50	0.20	0.50	0.25	0.50	0.00
i_2	1.00	0.50	0.75	0.50	0.80	0.50	0.75	0.50	1.00
i_3	0.00	0.50	0.25	0.50	0.20	0.50	0.25	0.50	0.00
i_4	1.00	0.50	0.75	0.50	0.80	0.50	0.75	0.50	1.00
i_5	0.00	0.50	0.25	0.50	0.20	0.50	0.25	0.50	0.00
i_6	0.00	0.00	0.25	0.25	0.40	0.25	0.50	0.50	1.00

From this example, it is found that when the word diversity is higher, the extent of decline of the indexing phase performance will be larger. It is because when the words are more distinct from each other, the concept headings converted from these words will also be more diverse. The chance will be higher to propose incorrect indexes unintentionally.

From this example and the previous example, it is found that although the relative proposing frequencies of correct indexes are reduced but in most cases, they are still higher than those of incorrect indexes. Therefore, the correct indexes can be selected out. But when the word diversity of documents

is getting larger and larger, the decline of indexing phase performance will become more serious that correct indexes cannot be selected.

Table 9 Index proposing frequencies suggested by using words and using concept headings found in the documents listed in the table 7

Doc. 0

indexes:	i_0	i_1	i_2	i_3	i_4	i_5	i_6
PFw:	0.22	0.00	0.33	0.00	0.33	0.00	0.11
PFc:	0.15	0.10	0.24	0.10	0.24	0.10	0.09

Correct indexes	(PFw-PFc)/PFw
i_0	0.318
i_2	0.273
i_4	0.273

Doc. 1

indexes:	i_0	i_1	i_2	i_3	i_4	i_5	i_6
PFw:	0.00	0.33	0.00	0.33	0.00	0.33	0.00
PFc:	0.10	0.14	0.20	0.14	0.20	0.14	0.10

Correct indexes	(PFw-PFc)/PFw
i_1	0.576
i_3	0.576
i_5	0.576

Doc. 2

indexes:	i_0	i_1	i_2	i_3	i_4	i_5	i_6
PFw:	0.11	0.00	0.33	0.00	0.33	0.00	0.22
PFc:	0.09	0.10	0.24	0.10	0.24	0.10	0.15

Correct indexes	(PFw-PFc)/PFw
i_2	0.273
i_4	0.273
i_6	0.318

On the whole, the word diversity of documents is a critical factor to determine whether it is worth using concept headings in the automatic indexing. When the word diversity is low, one can consider to use them.

3.9 Correctness prediction for proposed indexes

According to routine procedures of automatic indexing, there will be a certain number of indexes proposed for a non-indexed document, no matter the indexes are correct or not. For example, if many words of a non-indexed document are not yet covered in the training phase, the proposed indexes for this document will be inappropriate since these words do not have any association with indexes. Under such a situation, the user of this automatic indexing method should be informed that the index assignment of this document may not be correct. For an ideal automatic indexing method, the method itself should be able to estimate the degree of correctness of proposed indexes. In other words, the correctness prediction should be done automatically.

In the past researches, the method of predicting the correctness of proposed indexes is seldom mentioned or ignored. In this study, a simple method is introduced in order to predict the correctness of proposed indexes. If an index is suitable for a document, the proportion of words proposing this correct index should be large. Conversely, if an index is not suitable, the proportion of words proposing this incorrect index should be small. An index proposing rate (PR_j) of an index i_j in a document is defined to represent the proportion of words proposing an index.

$$PR_j = \frac{\text{no. of words proposing an index } i_j}{\text{total no. of words in a document}} \qquad (7)$$

If the PR_j is large, there will be higher chance that the index i_j is a correct one. Sometimes the proposing frequency of a false index may be very high. This index may be proposed by only few words which have large $P(i_j/w_i)$ values with this index. The advantage of using the index proposing rate is that this parameter value is independent of the $P(i_j/w_i)$ which may be incorrect due to statistical errors induced in the training phase (eg. low occurrence frequencies of words). The index proposing rate is only related to a non-index document's features such as presence of words and their frequencies in the document.

If the proposed indexes are with a high precision (ie. more indexes are correct), the average index proposing rate of these proposed indexes will become large. Conversely, if the proposed indexes are with a low precision (ie. few indexes are correct), the average index proposing rate of these proposed indexes will become small. It is because larger index proposing rates of correct indexes can lead to a larger average index proposing rate. Therefore, there will be a correlation between the average index proposing rate of proposed indexes and the precision of them. In other words, using the average index proposing rate of proposed indexes can predict the precision of them.

3.9.1 Example of using index proposing rate

The material used in the example for demonstrating the processes of training phase and the indexing phase will be used here to illustrate the use of index proposing rate.

The four documents are listed below.

```
Doc 1
Index: VITAMIN, CARBOHYDRATE
Text: Vitamin B and starch are rich in rice.

Doc 2
Index: CARBOHYDRATE, FAT
Text: Starch and fat are rich in peanut.

Doc 3
Index: FAT, PROTEIN
Text: Fat and protein are rich in meat.

Doc 4
Index: VITAMIN, PROTEIN
Text: Vitamin B and protein are rich in fish.
```

The statistical correlations between indexes and words found in these documents are shown on the table below.

	"starch"	"protein"	"fat"	"vitamin B"
CARBOHYDRATE	2/2	0/2	1/2	1/2
PROTEIN	0/2	2/2	1/2	1/2
FAT	1/2	1/2	2/2	0/2
VITAMIN	1/2	1/2	0/2	2/2

Now, the Doc 1 will be treated as if it is a non-indexed document. The words (underlined) of this document will be used to propose indexes. The proposing frequency of each index will be calculated as before. Moreover, the index proposing rate of each index will be calculated.

Index proposing rate of PROTEIN for Doc 1

$$= \frac{\text{no. of words proposing PROTEIN in Doc 1}}{\text{total words of Doc 1}}$$

$$= \frac{\text{no. of words (in Doc 1) having non-zero correlation with PROTEIN}}{\text{total words of Doc 1}}$$

$$= 1/2$$

Similarly, other index proposing rates of indexes for Doc 1 are calculated and shown as below.

Doc 1

Indexes	Proposing frequencies	Index proposing rate
CARBOHYDRATE	1.5	1.0
PROTEIN	0.5	0.5
FAT	0.5	0.5
VITAMIN	1.5	1.0

First, it is found that the proposing frequencies of correct indexes (ie. CARBOHYDRATE and VITAMIN) are higher than incorrect ones. Second, the index proposing rates of indexes are higher if the indexes are correct. Now, some different combination of proposed indexes will be attempted in order to observe the change of average index proposing rate. Following are these combinations and average index proposing rates of them.

Index combination	Average index proposing rate
CARBOHYDRATE + VITAMIN (2 correct)	1.000
CARBOHYDRATE + VITAMIN + FAT (2 correct + 1 incorrect)	0.833
CARBOHYDRATE + PROTEIN (1 correct + 1 incorrect)	0.750
CARBOHYDRATE + PROTEIN + FAT (1 correct + 2 incorrect)	0.667
PROTEIN + FAT (2 incorrect)	0.500

From this example, it is found that when a larger proportion of proposed indexes is correct (ie. precision of them is higher), the average index proposing rate will be larger. The average index proposing rate of a group of proposed indexes can be used to predict the correctness of this group of indexes.

3.10 Effect of subject matter on automatic indexing

In the following paragraphs, the effect of subject matter on the statistical approach of automatic indexing will be discussed. Because it is reasonably to believe that the conditions for a practical automatic indexing method are not identical for all disciplines. Different disciplines can have different effects on the performance of automatic indexing.

Rowbottom and Willett [16] used the approach same as that used by Pao [15] to perform automatic indexing. They attempted different disciplines: natural science, medicine, mathematics, social science, political science, humanities, and technology and engineering. Rowbottom and Willett have shown that the subject matter strongly affects indexing performance since scientific and technological extracts are generally assigned many more index terms than extracts from the social sciences and humanities. It is because there is smaller proportion of words occurring only once in the science papers when compared with other disciplines. There will hence be a greater number of terms above and below the transition point which will be selected using the Pao algorithm.

From their study result, it is found that in scientific and technological extracts, words will occur more frequently to reflect the main concepts of the documents. Therefore, scientific and technological documents are relatively

easy to be indexed correctly. Hamill and Zamora [6] developed an automatic document classification system for documents about chemistry while Maron [13] performed his experiment using documents about computer science. They had obtained satisfactory results to demonstrate the feasibility of applying the automatic indexing for scientific papers.

3.11 Comparison with other indexing methods

In these paragraphs, some features of the automatic indexing method described in this study will be compared with those of other indexing systems. First, differences among different statistical approaches of automatic indexing will be discussed. Then, differences between automatic indexing and manual indexing will be covered.

Among different statistical approaches of automatic indexing, the major difference is the control of indexing language. The indexing languages can be divided into two types based on the degree of control of indexing languages allowed to be used. These two types are controlled-language indexing and natural-language indexing. For controlled-language indexing, a list of index terms allowed to be used will be determined in advance. The indexers can only use the index terms found in this list. The index terms are usually arranged in a hierarchial order in the list. The Medical Subject Heading (MeSH) is a typical example of this kind of indexing language. For natural-language indexing, the terms (words or phrases) found in a document (ie. title, abstract or full text) will be selected and used as index terms for the document.

Both the controlled-language indexing and natural-language indexing have been attempted in automatic indexing by researchers. Hamill and Zamora [6] have developed an automatic classification systems for a certain type of

controlled-language indexing. For natural-language indexing, Goffman [5] and Pao [15] have developed a method to select some terms from a document text as indexes for the document. The indexing method proposed in this study is for controlled-language indexing.

For using natural-language in automatic indexing, the major technique used is to select some content-bearing words based on frequencies of words found in a document. As mentioned earlier, Goffman used the Zipf's law and the Booth's law to identify words that have medium occurrence frequencies. In his approach, assigning indexes to each document is performed regardless of other documents. The features (ie. frequencies of words) of a document completely determine the indexes assigned to the document. Thus, in this approach, the relations between words are emphasized. For example, co-occurrence of two words will be used as a hint to determine the relations between them.

For using controlled-language in automatic indexing, there are some differences. The major technique used is to calculate statistical correlations between indexes and words found in some indexed documents. Then, based on these index-word associations and the frequencies of words of a non-indexed document, the indexes are assigned. Thus, the index assignment of a document is partially dependent on the features of the non-indexed document and partially dependent on the features of indexed documents that have been used

to calculate index-word associations. In this case, the co-occurrences between indexes and words are emphasized. The statistical correlations between words and indexes are mainly used to determine the indexes.

The advantage of using controlled-language indexing is that the index terms assigned to documents can be kept in a higher degree of consistency. Also, using hierarchial relations between index terms can assist in the searching process. But using controlled-language requires the list of allowable index terms to be defined in advance. This step needs expertise. Moreover, the indexers require more effort to do the indexing. Also, the searchers need to consult this index list before constructing the query for searching information. For example, if one wants to search something about "gossypol acetate", he should use the MeSH term "Gossypol--Analogs and Derivatives" after he has referenced the list of allowable index terms. The advantages of using natural-language indexing are as follows. Less effort is required in the indexing stage. Sometimes, natural-language may reflect more closely the terms used by the searchers. For example, searching for documents about Chalets and this is not an index in the list of allowable index terms. But using natural-language has some drawbacks. First, the synonyms and variant words of natural language can lead to lower consistency of assigning indexes. Also, the hierarchial relationship and cross reference between indexes cannot be expected. Using broader or

narrower concepts relies heavily upon the knowledge and experience of the searchers.

Now, the differences between automatic indexing and manual indexing are discussed. The major difference between them is the coordination of indexes. There are two types of coordinations: precoordination and postcoordination. Index coordination is an indexing scheme that combines single index terms to create composite subject concepts (eg. the index terms EYE and SURGERY are combined to create concept eye surgery). The system allows the coordination of classes either before or during searching. In precoordination, the combination are made at the indexing stage by the indexers while in the postcoordination, the combinations are made at the searching stage by the searchers. For example, the index terms assigned to a document with the title "using mountain camping equipment in the environment with desert climate" are MOUNTAIN, CAMPING EQUIPMENT, DESERT and CLIMATE. In the precoordination, indexers will arrange and link these four indexes as MOUNTAIN-CAMPING EQUIPMENT and DESERT-CLIMATE while in the postcoordination, this arrangement is not managed by the indexers and the relations between these four indexes are not indicated in the indexing stage. The precoordination is better than the postcoordination because predefined coordination between indexes can reduce the chance of ambiguity caused by unclear relations between indexes. For example, if there

is another document with the title "using desert camping equipment in the environment with mountain climate", the indexes terms assigned with be same as those for the previous example (ie. "using mountain camping equipment in the environment with desert climate"). Without the precoordination for indexes, these indexes cannot be used to separate these two documents. In the search stage, the searcher may use AND operators to link these four indexes and these two documents will be retrieved altogether.

Since the precoordination requires the analysis of the relations between assigned indexes, it can be performed in the manual indexing but it is relatively difficult to be implemented by the automatic indexing of statistical approach. Typically, the automatic indexing can only propose some indexes but cannot construct the coordinations between them. Thus, the automatic indexing belongs to the scheme of postcoordination.

3.12 Proposal for applying Chinese medical knowledge

The contents of collected articles in the CMMRC database are about the research of the Chinese herbs. In these documents, herb names are already written in both scientific name and Chinese name simultaneously. For example, the herb name Dang Gui in the text is described by a corresponding scientific name (Latin name) *Angelica sinensis*. Now in the current database system of CMMRC, it is already possible to use a scientific name of a herb to retrieve all documents mentioning this herb. For example, if one uses *Angelica sinensis* as a search term, all documents that have mentioned Dang Gui will be retrieved. Because each scientific name is unique for a certain herb. In the future, when indexes are added for each document, the scientific name and Chinese name of a herb can be treated as index terms.

However, although the data in the database are about the Chinese herbs, in most of these articles, herbs are described to be studied by modern scientific methods. There is little stored information about traditional Chinese medical knowledge applied on these herbs. It will be more effective if one can apply some traditional Chinese medical knowledge in the index assignment and the information retrieval in the CMMRC database. In fact, the application of Chinese medical knowledge is being accentuated in the research of Chinese herb in recent years. At least, this is the case in the CMMRC.

According to theory of traditional Chinese pharmacology [14], each Chinese herb has fixed combination of properties and flavours. Various properties and flavours of herbal medicines exert different effects. There are four properties of herbal medicines, ie. cold, heat, warm and cool. In general, the herbal medicines with warm and heat properties are prescribed for cold-syndrome (eg. aversion to cold, cold limbs, pale tongue, slow pulse, etc) and those with cool and cold properties for heat-syndrome (eg. fever, thirst, deep-colored urine, red tongue, rapid pulse, etc). The herbal medicines are grouped under five flavours, ie. acridness, sweetness, sourness, bitterness, and saltiness, which exert different effects. Generally speaking, acridness serves to expel and to activate; sweetness, to invigorate, to regulate and to moderate; sourness, to astringe and to preserve; bitterness, to lower, to release and to dry; saltiness, to soften and to purge.

Every herbal medicine possesses a specific property and flavour of varying degree. It is combination of both that constitutes the overall action of individual medicine. On the other hand, herbal medicines may has various pharmacological actions. For example, Ginseng has tonic action while Ma Huang can induce sweating to expel the exogenous evils from the body surface. The pharmacological actions of a herbal medicine is always not one but many. For example, Niu Huang is not only a phlegm-eliminating agent but also a heat-clearing one. The herbal medicines which have common pharmacological actions, somehow, can be used for treating same syndrome. For instance, Tu Fu

Ling and Jin Yin Hua have heat and toxin clearing action. They are prescribed for heat-syndrome such as fever, thirst, deep-colored urine, red tongue, rapid pulse, etc.

Since herbs' properties and flavours are essential features which are not yet covered in the current database system, the property-flavour combination of each herb can be added and used as an index term for the herb in the CMMRC database. Now, the property-flavour combinations of many herbs are well-documented in many Chinese medical articles. The attachment of this kind of information in the database is just a clerical work requiring a little expertise.

Chapter four

Simulations of automatic index generation

In order to verify the procedures used in the automatic index generation and to study the factors that will affect the performance of these procedures, a series of simulations of these procedures have been performed. In these simulations, imaginary data will be used since the factors affecting the performance can be readily controlled. There are several simulations performed to test: (1) training phase performance, (2) indexing phase performance, (3) performance of using concept headings, and (4) performance of using index proposing rate to predict the correctness of proposed indexes.

4.1 Training phase simulations

In order to verify the essential procedures of the training phase, simulations are performed to test the last two processes of training phase (ie. the process of calculating associations between indexes and words, and the process of discarding false associations). As mentioned before, the performance of the training phase is dependent on the size of training document set and the word diversity of each document. These two factors will be taken into the consideration in the simulations.

Two types of simulations will be performed. First type is designed to test whether correct associations between indexes and words can be constructed by the procedures of the training phase. This type of simulation will be performed with two different controls of word diversity respectively. One simulation will be performed when the word diversity is not controlled (ie. in random manner) while another will be performed when the word diversity is controlled. Another type of simulation is to test whether the use of $(\Sigma\triangle R_{ij})/F_i$ can successfully assist in discarding false associations.

4.1.1 Simulation of association calculation (word diversity uncontrolled)

Following are some assumptions of this simulation whose objective is to test whether the correct associations between indexes and words can be searched and extracted out from training documents when words can be randomly grouped in each training document (ie. word diversity is not controlled).

Assumption

(1) It is assumed that there is a small database system of documents of free text. The number of predefined indexes allowed to be used in this simulation is 50. These indexes are represented by symbolic code i_1, i_2, i_3, ... i_{50}.

In addition, the number of predefined words allowed to be used is 50. These words are represented by w_1, w_2, w_3, ... w_{50}.

(2) It is predefined that for each index, three words are truly associated with it. The association strengths of these words are assumed to be identical. These predefined associations are randomly generated in advance. Because of this random manner, each word can be associated with one or more indexes.

(3) A number of training documents are generated for use in the training phase. For each document, it is predefined to be made up of three words and indexed by three indexes. Words for each document are randomly assigned. Each word will be indexed by an index respectively based on predefined associations made in (2) above. If a word is predefined to be associated with two or more indexes, one of these indexes will be randomly selected to be assigned. Also, it is assumed that words in documents are already processed by the stopword elimination and the word standardization.

The aim of this design is to distribute these predefined associations between indexes and words into training documents randomly. Then, one can check whether these predefined associations can be rebuilt. The performance can be measured by comparing rebuilt associations with predefined ones. Since in each training document, the combination of words is randomly determined, the diversity of words found in these documents is high.

Performance measurement

The training phase performance can be practically measured by precisions of rebuilt associations and recalls of predefined associations. Precision and recall of associations between an index and words are defined as follows.

$$\text{Precision} = \frac{\text{number of correct words proposed}}{\text{total number of words proposed for rebuilt associations of an index}} \quad (8)$$

$$\text{Recall} = \frac{\text{number of correct words proposed}}{\text{total number of words in predefined associations of an index}} \quad (9)$$

Both the precision and recall are parameters conventionally used to represent the performance of information retrieval system in many past researches.

After the training phase, each index will be associated with candidate words by different $P(i_j/w_i)$ values. Candidate words with highest $P(i_j/w_i)$ values will be selected for the index. Theoretically, if more words are selected, the precision will be reduced while the recall will be increased. It is because when more words are selected to increase the recall, there will be more incorrect words selected simultaneously to reduce the precision.

For example, assume predefined associations between an index i_1 and words are as follows.

index	associated words
i_1	w_1, w_3, w_5

After the training phase, the rebuilt associations are as follows.

i_1 <-- w_1, w_2, w_3, w_4, w_5, w_6 .... (sorted by $P(i_j/w_i)$ with descending order)

If first three words are chosen to calculate, both the precision and recall will be 2/3. But if first six words are chosen, precision will be reduced to 3/6 and recall is increased to 3/3.

Theoretically, the higher the precision and recall values, the better the training phase performance will be. In order to evaluate the overall performance of the training phase, the average precision is used and defined to be an average of all (ie. 50) rebuilt associations' precisions. Similarly, the average recall is used and defined to be an average of all predefined associations' recalls.

Procedures

Training documents in this simulation will be processed by the procedure of calculating associations between indexes and words. This means that associations between each index and corresponding words will be constructed. Then, the average precision and the average recall of these rebuilt associations will be calculated. A set of simulations have been performed to calculate the

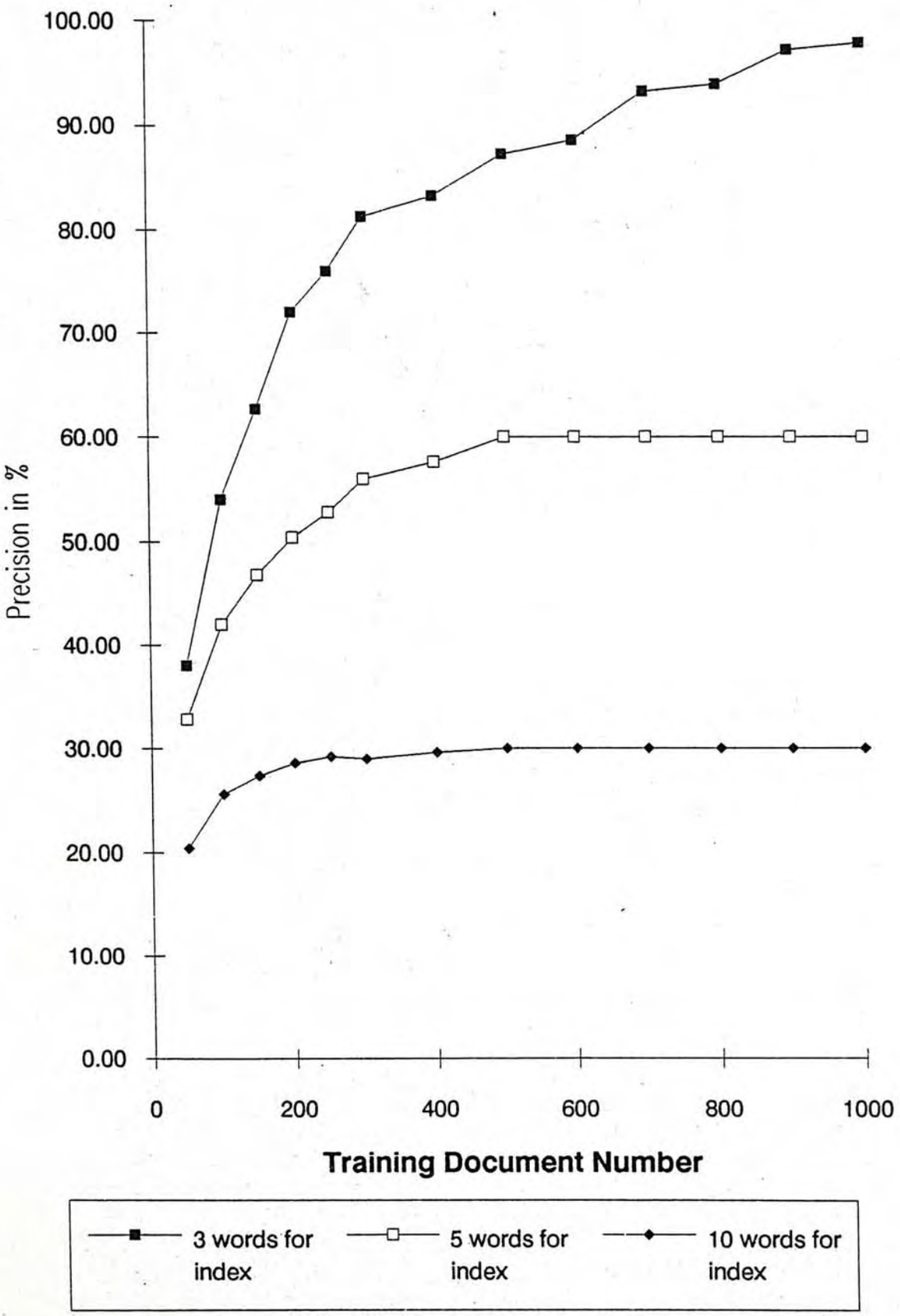
average precision and the average recall under different training document number and different number of words (with highest $P(i_j/w_i)$ values) selected for each rebuilt association.

Result and analysis

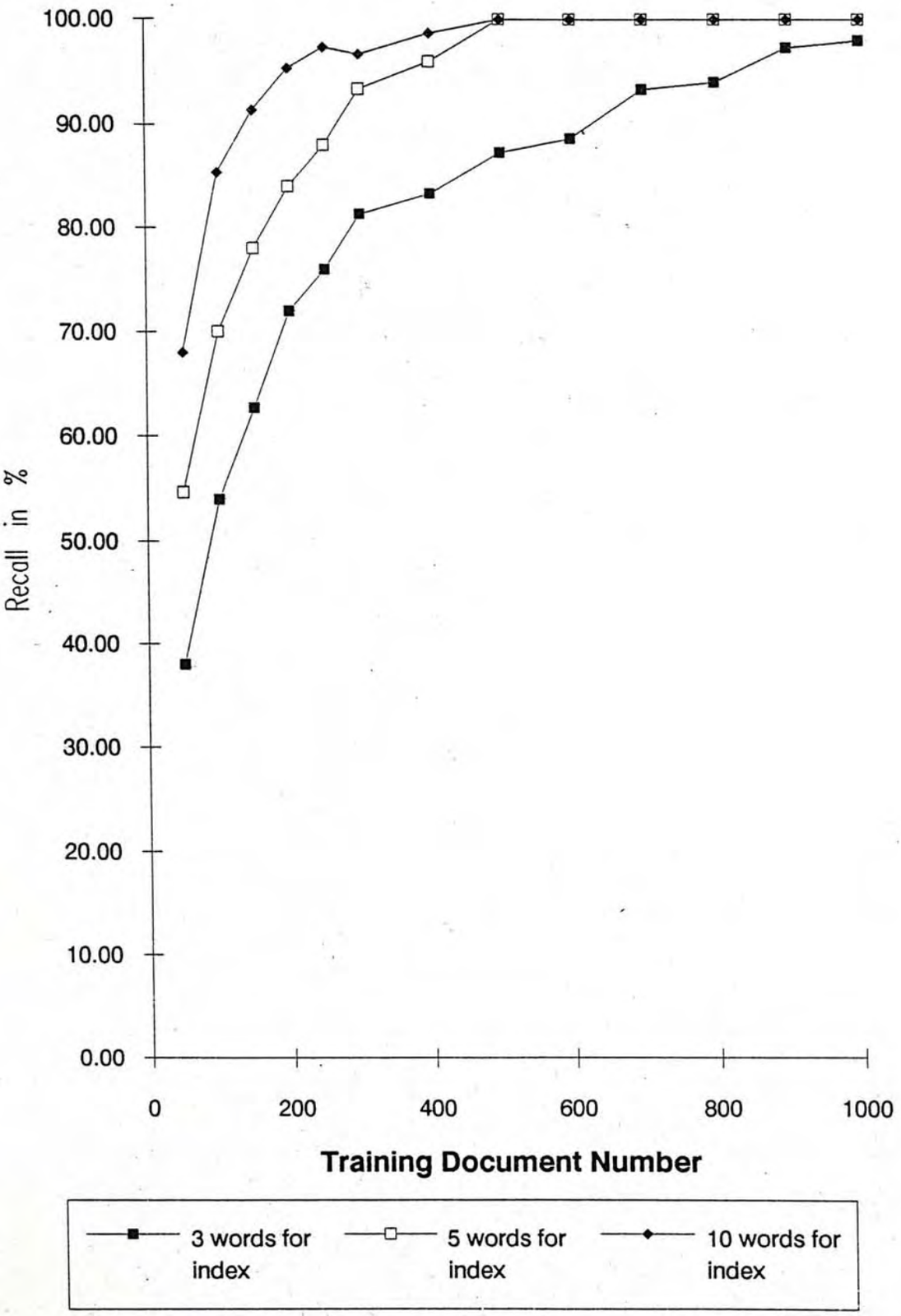
Results are presented in the graph 1 and 2 on next two pages. First, it is discovered that the average precision and the average recall are increasing with training document number. When the training document number is 1,000 and only three words are selected for each index, 98% of predefined associations can be rebuilt successfully. Thus, this result shows that provided that there are sufficient training documents, correct associations between indexes and words can be successfully constructed.

In these simulations, when the training document number is larger, there will be higher chance for predefined associations to be rebuilt correctly. As mentioned earlier, the number of training documents used in the training phase should be large enough to cover sufficient topics of contents and reduce statistical errors caused by low occurrence frequencies of words in order to achieve a reliable training result. Thus, this result agrees with the predicted effect of training document number on the performance of the training phase. If it is possible, one should use more training documents in the training phase.

Graph 1
Change of precision with training document number



Graph 2
Change of recall with
training document number



Second, it is found that values of precision and recall are level off at the same time in the graphs. It is because when the recall is 100%, all predefined (correct) words associated with each index have been successfully proposed. Therefore, any increase in the training document number cannot further increase the recall value. On the other hand, when recall is 100%, the number of correct words proposed for each association has reached its maximum limit, the precision value is, therefore, level off at a certain value.

Third, it is found that before values of precision and recall are level off, for a particular number of training documents, an increase in number of words selected for rebuilt associations can cause the precision to be reduced and the recall to be increased. This result agrees with theoretical relationship between the precision and the recall.

4.1.2 Simulation of association calculation (word diversity controlled)

In the previous simulation, in each training document, each word is indexed by one index. Because of the random grouping combination of words in a document, each index may be only related to one word and unrelated to others. Therefore, the random manner of word assignments which, in turn, leads to random combination of indexes can cause many false associations between words and indexes. It is because the word diversity is high in these documents. Also, in this circumstance, many training documents are required in order to develop correct associations between words and indexes. In the previous simulation, one needs about 1,000 documents to attain a satisfactory performance.

In order to verify the effect of the word diversity, another simulation is performed to test the performance of training phase when the word diversity of each document is controlled and kept in a relatively low level.

Assumption

The assumptions and procedures of this simulation is similar to those of the previous one. But the differences in this simulation are that only 50 training documents will be used and each word is predefined to be associated with exactly three indexes. Word diversity of each document will be controlled.

Different performances due to different word diversities will be compared. Following is the mathematical definition of word diversity used in this simulation.

It is assumed that if two words are identical, these two words will be associated with common indexes. Thus, one can use this feature to measure the similarity of two words and represent the word diversity of a document. The similarity of words w_a and w_b, S^i_{ab}, will be defined as follows.

$$S^i_{ab} = \frac{n_{ab}}{\sqrt{n_a \times n_b}} \qquad (10)$$

where n_{ab} is number of indexes commonly associated by w_a and w_b while n_a and n_b are number of indexes associated by w_a and w_b respectively.

The above equation is a simplified version of the cosine correlation. In this simulation, each predefined index-word association is assumed to have equal importance. One only considers whether a certain word is associated with a certain index. In other words, weight of a certain index-word association is one or zero. Thus, one only needs to count how many indexes are associated with a word and how many indexes are commonly associated with two words.

Assume in a document, there are N words. The average word similarity, AWS, which reflects the word diversity of a document has been defined (in the chapter three) to be as follows.

$$AWS = \frac{\sum_{i=1}^{N} \frac{\sum_{j=1, i \neq j}^{N} S_{ij}}{N-1}}{N} \qquad (5)$$

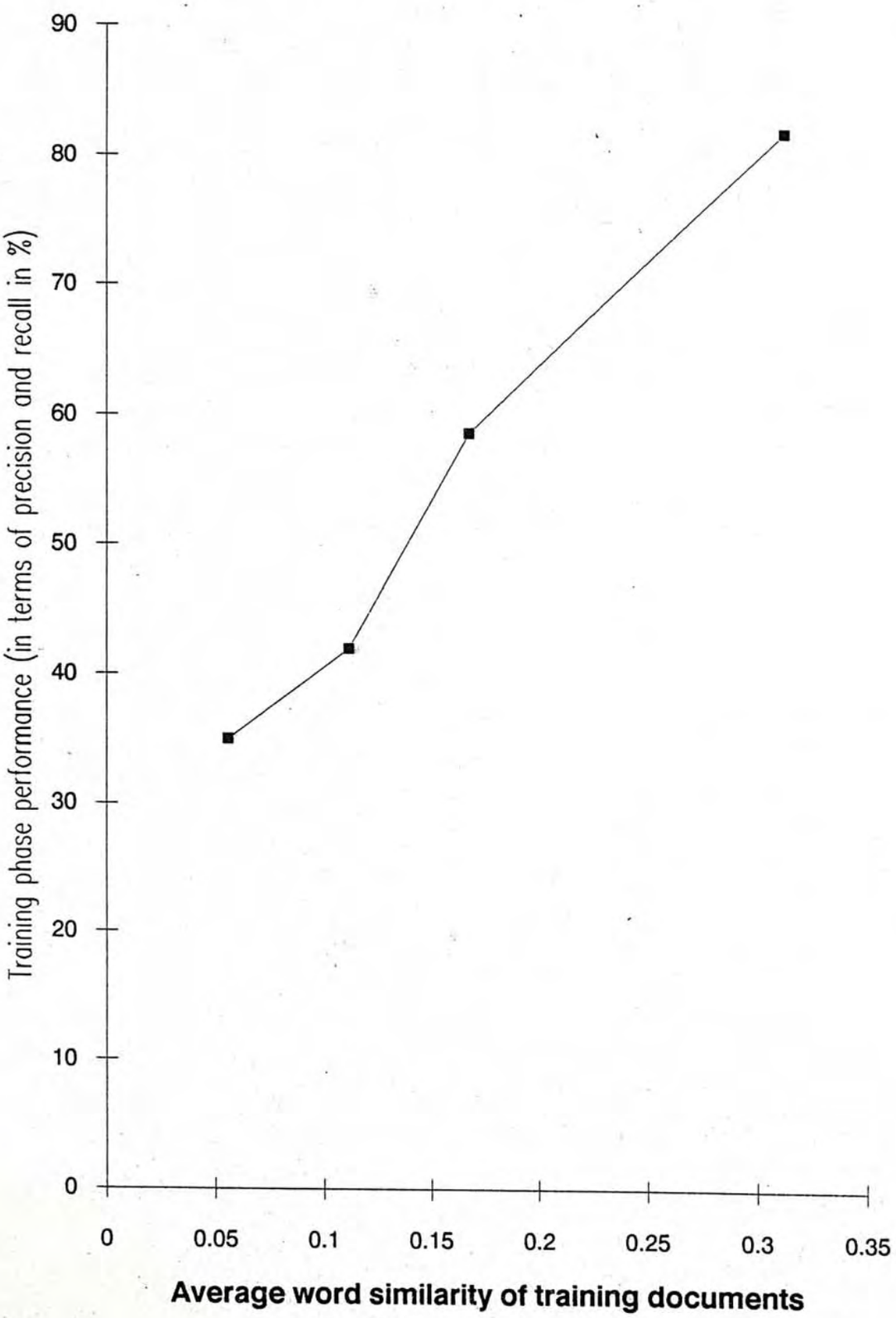
According to the above definition, the word diversity of a document will be increased with the decrease of AWS. In this simulation, the S_{ij} in the equation 5 is replaced by S^{i}_{ij}.

Result and analysis

The result of this simulation is shown on graph 3 on next page. In this simulation, the number of words selected for each index is three (ie. original number of words in each predefined association with an index). Thus, the values of precision and recall will be identical and represented by one line in the graph.

It is found that when the word diversity is decreasing (ie. AWS is increasing), the performance (in terms of precision and recall) of the training phase is being improved. This result is consistent with the predicted effect of word diversity of training documents. Thus, when the related words are

Graph 3
Change of training phase performance with average word similarity of training documents



grouped together to be indexed by a set of related indexes, the performance of the training phase can get better.

4.1.3 Simulation of discarding false associations

The last process in the training phase is to discard false associations. A simulation has been performed to test whether the value of $(\Sigma\Delta R_{ij})/F_i$ can effectively assist in determining which words are truly associated with an index.

Assumption

Assumptions of this simulation are as follows.

(1) The number of predefined indexes allowed to be used is 50 and the number of predefined words allowed to be used is also 50.

(2) In this simulation, it is predefined that for each index, five words are truly associated with it. The association strengths of these words are assumed to be identical. These predefined associations between indexes and words are randomly generated in advance. Because of this random manner, each word can be associated with one or more indexes.

(3) A number of training documents are generated for use in the training phase. For each document, it is predefined to be made up of five words and indexed by five indexes. Words for each document are randomly assigned. One index is assigned to index each of these words based on predefined associations made in (2) above. If a word is predefined to be associated with

two or more indexes, one of these indexes will be randomly selected to be assigned. Also, it is assumed that words in documents are already processed by the stopword elimination and the word standardization. Note that in this simulation, the word diversity of each training document is not controlled.

Procedures

In this simulation, there will be several training phases. In the initial training phase, 500 training documents will be used. Fifty training documents will be increased for each successive training phase. In the final training phase, there will be 1,500 training documents. For each training phase, the ΔR_{ij} will be calculated and recorded for each word in an association with each index. After all training phases have been proceeded, value of $(\Sigma\Delta R_{ij})/F_i$ will be calculated for each word in an association with each index.

Performance measurement

Since the value of $(\Sigma\Delta R_{ij})/F_i$ is used to distinguish truly associated words from trivial words. The performance of this method can be measured by comparison of $(\Sigma\Delta R_{ij})/F_i$ values of candidate words associated with each index. As mentioned before, in this simulation, five words are predefined to be associated with each index. If this method is successful, there should be an obvious change in the $(\Sigma\Delta R_{ij})/F_i$ value between word of 5th rank and word of

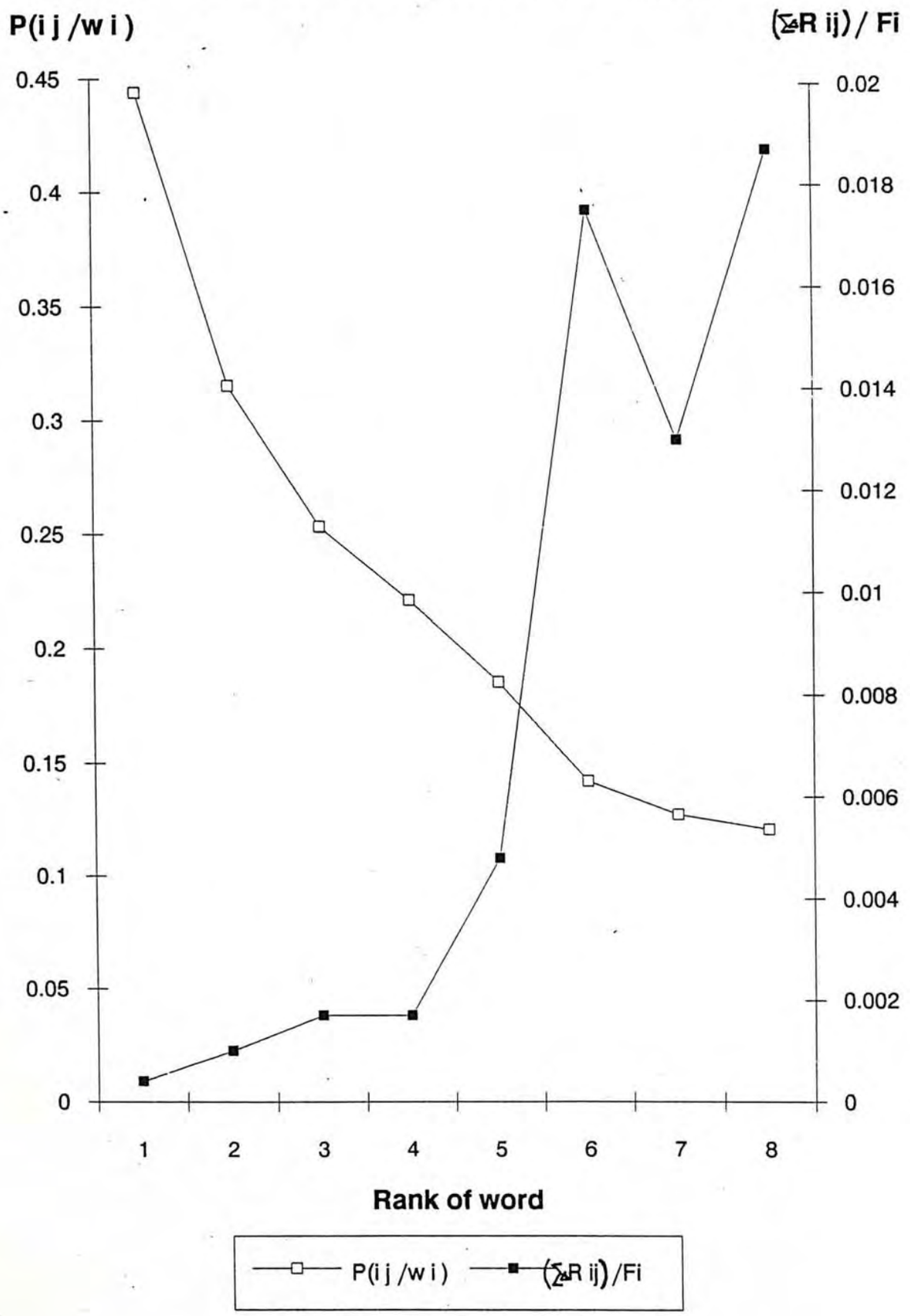
6th rank proposed by the final training phase. Thus, the overall performance can be observed by comparing the average $(\Sigma\Delta R_{ij})/F_i$ value of all words of same rank for the first few ranks. For clearly demonstrating the change in $(\Sigma\Delta R_{ij})/F_i$ value, the average $(\Sigma\Delta R_{ij})/F_i$ value of first eight ranks are checked.

Result and analysis

Average $(\Sigma\Delta R_{ij})/F_i$ values and average (Pi_j/w_i) values of first eight ranks in all associations proposed by the final training phase are shown on the graph 4 on next page.

From the result, it is found that there is, indeed, an obvious change in average $(\Sigma\Delta R_{ij})/F_i$ value between word of 5th rank and word of 6th rank. Also, it is discovered that average $(\Sigma\Delta R_{ij})/F_i$ values of first five ranks (words truly associated with indexes) are relatively small. Thus, the result agrees with the prediction that when the training document number is increased, rank changes of words truly associated with an index will be comparatively small. Therefore, the result shows that the $(\Sigma\Delta R_{ij})/F_i$ value can be used to determine which words truly associated with an index. On the other hand, the values of (Pi_j/w_i) only decrease gradually without obvious change to distinguish correct associated words from incorrect ones.

Graph 4
Change of average (∑R ij)/ Fi
and average P(i j / w i) with word ranks



Furthermore, it is found that when the rank is higher (ie. $P(i_j/w_i)$ value is larger), the $(\Sigma\Delta R_{ij})/F_i$ value will be smaller. This relationship between $P(i_j/w_i)$ and $(\Sigma\Delta R_{ij})/F_i$ can reflect the fact that words with higher probability (ie. higher $P(i_j/w_i)$) to be associated with an index are ones which are comparatively stable in their ranks.

Intermediate results when document number is increasing

In the previous simulation, one only considers the situation in which there are enough training documents used to perform sufficient training phases. Now, the situation in which there are fewer training documents will be considered. When the number of training documents is not enough, words will have fewer occurrence frequencies leading to lower chance of changing word ranks according to $P(i_j/w_i)$ values calculated after each training phase. Thus, rank changes of correct words and those of incorrect words will not be obviously different from each other.

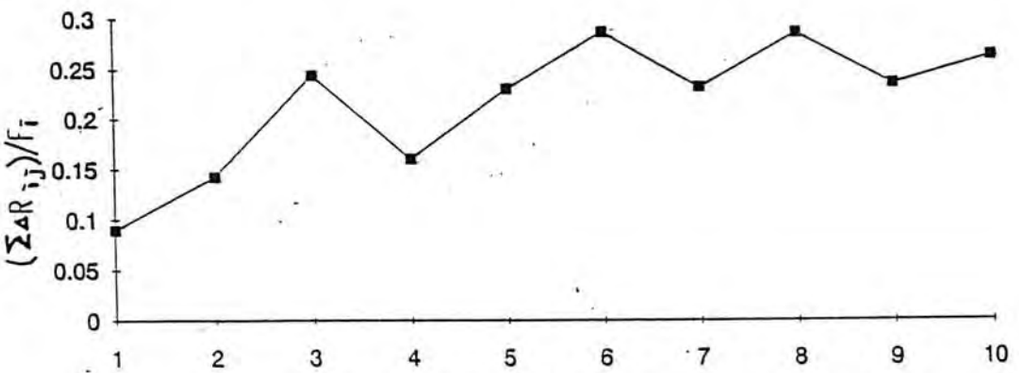
Now, there is another simulation. In this simulation, the initial training phase will have 50 training documents. For each successive training phase, 50 documents will be appended. For each index, the $(\Sigma\Delta R_{ij})/F_i$ values of first ten word ranks will be calculated after three particular training phases with three different document number: 300, 600 and 1,500 documents. It is expected that when there are more training documents, it is easier to detect the correct index-

word associations according to the $(\Sigma\Delta R_{ij})/F_i$ values. Like the previous simulation, each index is predefined to be associated with five words. Thus, if the performance is satisfactory, the change of $(\Sigma\Delta R_{ij})/F_i$ value will be obvious between word of 5th rank and that of 6th rank. Other assumptions used in this simulation are same as those used in the previous one.

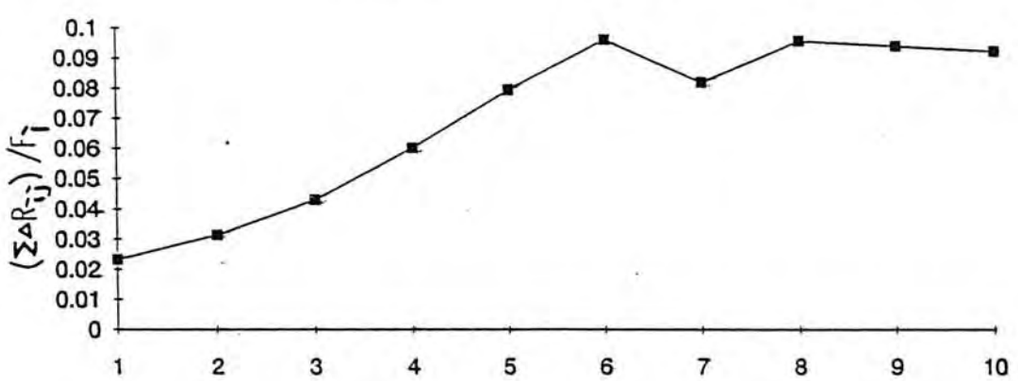
In the graph 5 on the next page, there are three different graphs illustrating the $(\Sigma\Delta R_{ij})/F_i$ values calculated under three different situations. When there are only 300 training documents, the curve has a zigzag shape. It is difficult to use these values to distinguish the correct associations from the incorrect ones since there is no obvious change of $(\Sigma\Delta R_{ij})/F_i$ value between two successive ranks. But when there are 600 training documents, the first several ranks have $(\Sigma\Delta R_{ij})/F_i$ values gradually increasing. Then, the curve is, more or less, level off. At least, this pattern of the curve is clear that the first few ranks with increasing $(\Sigma\Delta R_{ij})/F_i$ values are different from those with high and similar values. Finally, when there are 1,500 training documents, the first five ranks' $(\Sigma\Delta R_{ij})/F_i$ values have been decreased to a level obviously lower than other ranks' values. The words of first few ranks with low $(\Sigma\Delta R_{ij})/F_i$ values can easily be identified to be correct words that should be associated with the indexes. When the training documents are increasing from 600 to 1,500, the correct associations between indexes and words have been being developed and, thus, the correct words will have lower chances to change their ranks. In other words, their ranks are become stable when documents are increased.

Graph 5

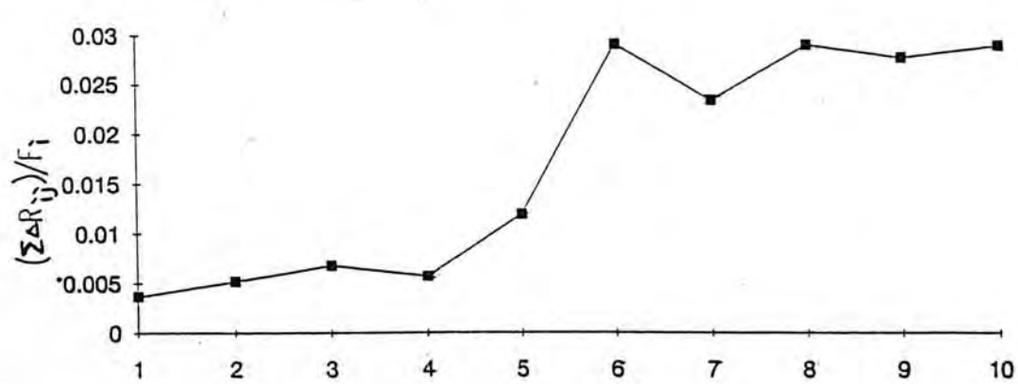
Change of $(\Sigma \Delta R_{ij})/F_i$ with word rank
(Final training phase with 300 documents
Performance = 74.0%)



Change of $(\Sigma \Delta R_{ij})/F_i$ with word rank
(Final training phase with 600 documents
Performance = 89.6%)



Change of $(\Sigma \Delta R_{ij})/F_i$ with word rank
(Final training phase with 1,500 documents
Performance = 97.6%)



From these three graphs, it is found that if a word is correctly associated with an index, there will have lower chance for this word to change its rank when the training documents are increasing. Conversely, if a word is not truly associated with an index, the rank change will be relatively serious when the training documents are increased.

4.2 Indexing phase simulation

The aim of this simulation is to test whether the indexing phase procedures can work properly to assign suitable indexes to a document. In this simulation, the training phase will be performed, in advance, to establish correct associations between words and indexes. Then, the training documents will be treated as if they are non-indexed documents. The words found in the documents will be used to propose the indexes based on the index-word associations calculated in the training phase. The performance of the indexing phase can be evaluated by comparing the proposed indexes with the original ones. The word diversity of the documents will be taken into consideration in the simulation since it is an important factor affecting the performance of this phase.

Assumption

In this simulation, the training phase will be accomplished in advance. The assumptions used in the training phase will be similar to those used in the training phase simulations mentioned earlier. These assumptions are as follows.

(1) The number of predefined indexes allowed to be used in this simulation is 50. In addition, the number of predefined words allowed to be used is 50.

(2) It is predefined that for each index, five words are truly associated with it. The association strengths of these words are assumed to be identical. Likewise, each word is predefined to be associated with five indexes.

(3) A number of training documents will be used in the training phase. For each document, it is predefined to be made up of five words and indexed by five indexes. It is assumed that words in documents are already processed by the stopword elimination and the word standardization. The grouping combination of words in these documents will be controlled in order to alter the word diversity intensionally.

Procedures

In this simulation, what are mainly investigated are the performance of the indexing phase and the factors affecting it. Therefore, before the indexing phase, the training phase will be deliberately controlled to make it rebuild the original predefined associations (ie. training phase precision and recall = 100% when first five words with largest association strengths are selected to calculate the precision and recall). Then, words of these documents will be used to propose indexes which will be compared with the original ones in order to calculate the performance.

In this simulation, the word diversity of each document will be controlled and different word diversity will be attempted in order to test its effect on the indexing phase performance.

Performance measurement

The performance measurement is accomplished by comparing the proposed indexes with original predefined ones. Precision and recall can be used to measure the performance of the indexing phase. In the indexing phase simulation, they are defined as follows.

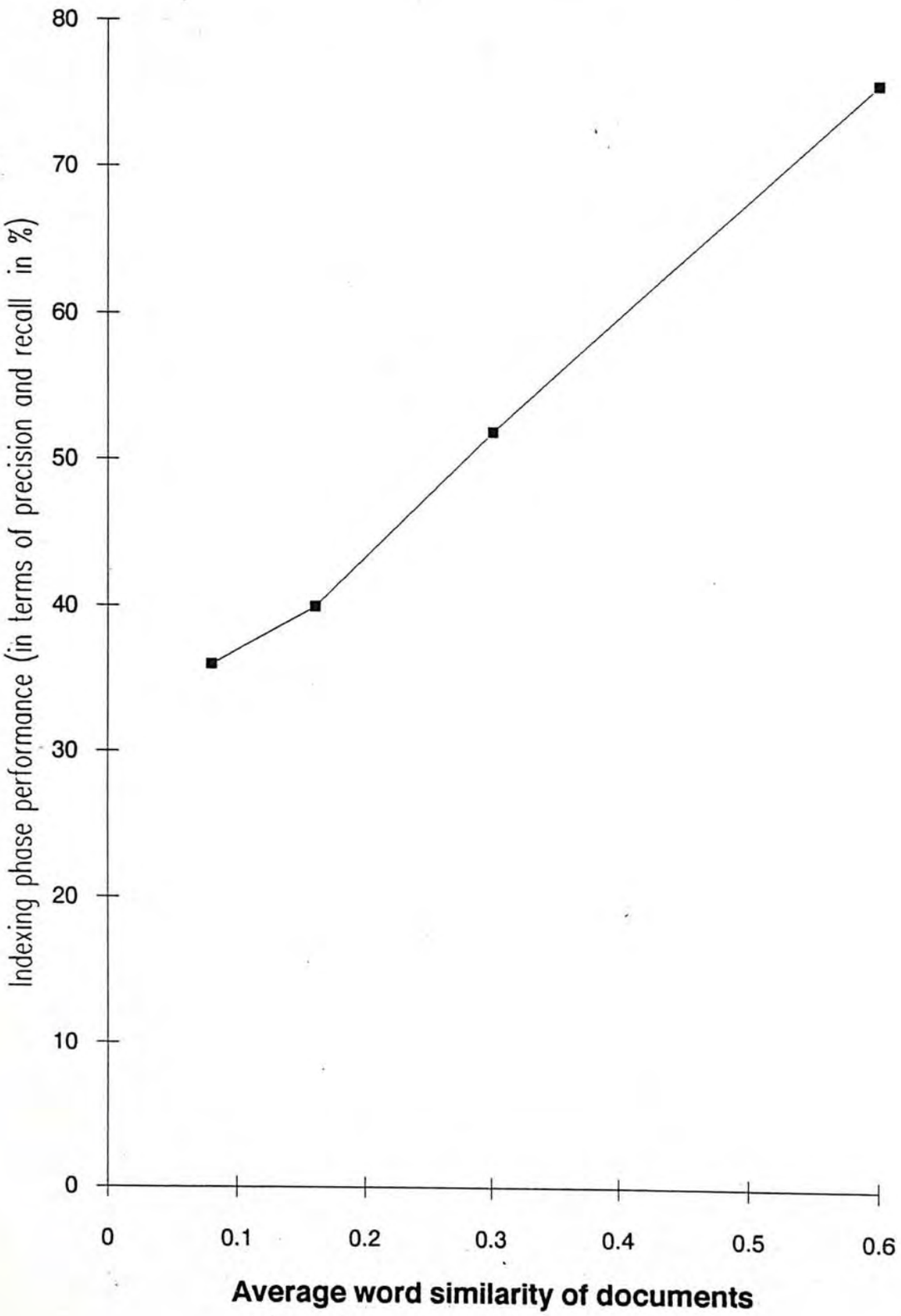
$$\text{Precision} = \frac{\text{number of correct indexes proposed}}{\text{total number of indexes proposed for a document}} \quad (11)$$

$$\text{Recall} = \frac{\text{number of correct indexes proposed}}{\text{total number of original indexes in a document}} \quad (12)$$

Result and analysis

The result of this simulation is shown on the graph 6 on next page. In this simulation, the number of words selected for each index is five (ie. original number of words predefined to be associated with an index). Thus, the values of precision and recall will be identical and represented by one line in the graph.

Graph 6
Change of indexing phase performance with average word similarity of documents



It is found that the performance of the indexing phase is related to the word diversity. The performance is becoming better when the AWS value is increasing (ie. the word diversity is decreasing). This result is consistent with the predicted effect of word diversity that when the words in a document are closely related to each other, the chance to get correct indexes will be higher. Because these words will concentrate on proposing some common indexes.

4.3 Simulation of using concept headings

The main aim of this simulation is to verify (1) the usefulness of concept headings and (2) the effect of word diversity on the indexing phase performance after concept headings are adopted.

As mentioned earlier, when the word diversity is higher than a certain level, the index phase performance will be affected significantly. A simulation is performed to test how the indexing phase performance will be affected if the concept headings are used. In this simulation, the indexing phase performance obtained by using concept headings will be compared with that obtained by using words.

Assumption

The assumptions used in this simulation will be identical to those used in the indexing phase simulation. But in this simulation, each word will be predefined to be represented by five concept headings. Each concept heading is predefined to be shared by five words. The number of concept headings allowed to be used in this simulation is 50.

Procedures

In this simulation, the same procedures will be performed for two different conditions. One is for using concept headings to calculate index-concept-heading associations and to propose indexes while another is for using words to calculate index-word associations and to propose indexes. Then, the differences between these two performances can be compared.

The concept headings found in each document will be utilized to perform the training phase to develop associations between indexes and concept headings. Then, these documents will be treated as non-indexed ones. The concept headings found in the training documents will be used to propose indexes which will be compared with the original ones. The definitions of precision and recall used in the indexing phase simulation will be used again in this simulation to measure the performance of the indexing phase.

Since the word diversity will affect the performance, this factor will be taken into consideration in the simulation. In order to evaluate this effect, different indexing phase performances due to different word diversities are compared.

In this simulation, the similarity of two words will be defined in terms of concept headings. If two words are exactly the same, they will share identical

concept headings. This feature can be used to measure the similarity of two words. The similarity of two words w_a and w_b, S^c_{ab}, has been defined (in the chapter three) to be as follows.

$$S^c_{ab} = \frac{c_{ab}}{\sqrt{c_a \times c_b}} \qquad (6)$$

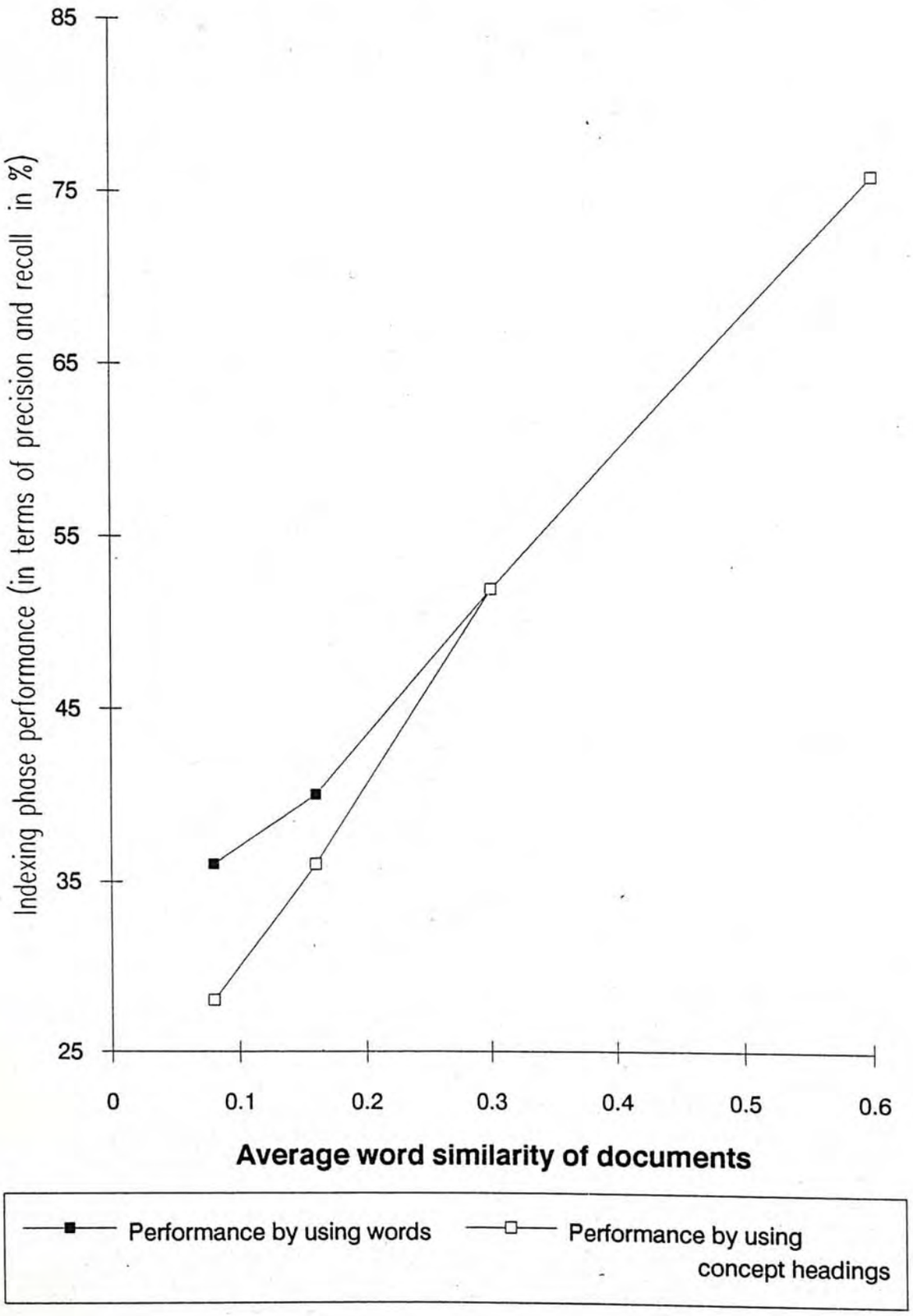
where c_{ab} is number of concept headings commonly shared by w_a and w_b while c_a and c_b are number of concept headings shared by w_a and w_b respectively.

In order to be consistent with the earlier definition of similarity S^i_{ab} (which is defined in terms of indexes rather than concept headings) in an equation 10, the combination of words and indexes in each document will be adjusted in order to make both similarity functions S^c_{ab} and S^i_{ab} return a common value for a document. The average word similarity, AWS, reflecting a word diversity of a document will be used as before (see equation 5).

Result and analysis

The result of this simulation is shown on the graph 7 on next page. In this simulation, the number of words selected for each index is five (ie. the original number of words predefined for each index). There are two lines on the graph: one for performance obtained by using words while another for performance obtained by using concept heading. It is found that when the word

Graph 7
Change of indexing phase performance with average word similarity of documents



diversity is increasing, the performance obtained by using words and that obtained by using concept headings are becoming poor. When the AWS is relatively high (>0.3), the performance obtained by using words and that obtained by using concept headings are same. But when the AWS is relatively low, the difference between two performances is apparent that the performance obtained by using concept headings is much poorer. Thus, this result is consistent with the predicted effect of using concept headings in the automatic indexing.

The result shows that only when the word diversity is rather high, the performance of indexing phase will be affected significantly. Otherwise, using concept headings will give same or similar performance compared with using words. Therefore, the advantages of using concept headings can compensate for the disadvantage of indexing phase performance decline which only occurs in documents with very high word diversity.

4.4 Simulation for testing performance of predicting index correctness

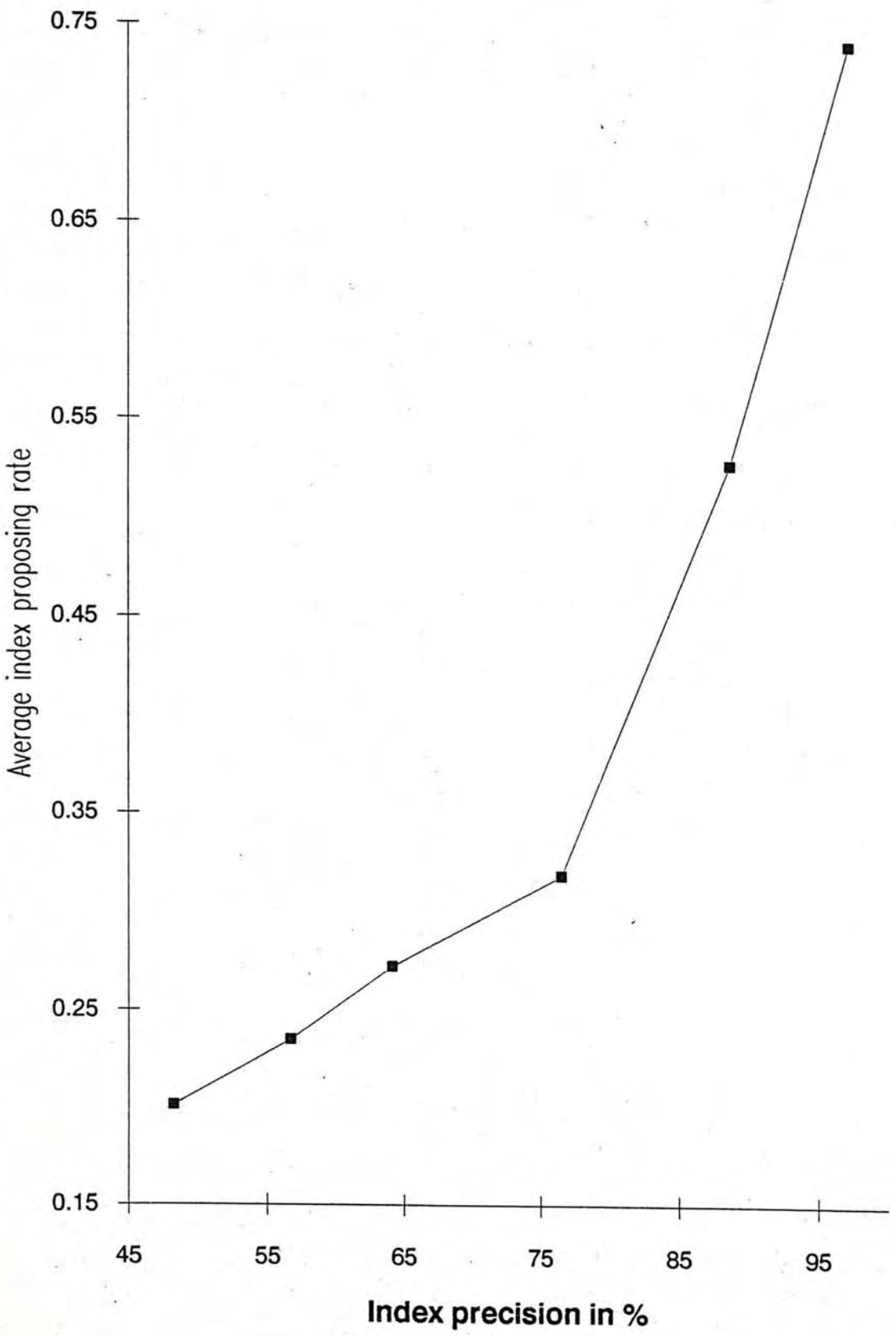
As mentioned before, the index proposing rate can predict the correctness of the proposed indexes. The aim of this simulation is to test whether there is a correlation between the index proposing rate and the correctness of proposed indexes and whether this feature can be used to predict correctness of proposed indexes.

The procedures of this simulation will be similar to those of the indexing phase simulation. In this simulation, each index is predefined to be associated with four words. Each document contains four words and is indexed by four indexes. In this simulation, different precision of proposed indexes will be attempted in order to examine whether the average index proposing rate of these proposed indexes will be changed by different precision values. The objective is to test if there is any correlation between average index proposing rate of proposed indexes and precision of them.

Result and analysis

The result is shown on the graph 8 on next page. In the simulation, for each document, four indexes with the highest proposing frequencies will be selected to calculate the precision using the equation 11. It is found that when the precision of proposed indexes gets higher, the average index proposing rate

Chapter 8
Change of index proposing rate with index precision



of these indexes also becomes higher. The result shows that there is correlation between the indexing phase performance and the average index proposing rate. When more proposed indexes are correct, the average index proposing rates of these proposed indexes will get higher since correct indexes will have higher index proposing rates. From this result, it is shown that the index proposing rate can be used to predict the correctness of proposed indexes. When the average index proposing rate of certain proposed indexes is higher, the precision of them will be larger.

4.5 Summary

In a series of simulations, a number of important aspects of automatic indexing are studied. The results of these simulations are summarized as follows.

First, it is found that the size of training document set is an important factor affecting the performance of training phase. The training phase performance can get improved if more training documents are used to calculate the index-word associations. It is because more training documents means more topics covered and fewer statistical errors caused by low occurrence frequencies of words and indexes found in them.

Second, the rank change of a word (according to $P(i_j/w_i)$ calculated each time when more training documents are used) in an association with an index can be used to distinguish the correct index-word associations from incorrect ones. It is found that a correct word will have relatively low chance to change its rank in the association with an index when training documents are increased.

Third, the performances of training and indexing phases are related to the word diversities of documents. When the word diversity is low, the performances of these two phases can get improved. For the training phase, low word diversity can reduce the number of false associations. For the

indexing phase, low word diversity can increase the proposing frequencies of correct indexes which can, then, be identified easily.

Fourth, it is found that the use of concept headings (ie. semantic representation of natural language terms) is feasible in the statistical approach of automatic indexing providing that the word diversities of documents are not rather high.

Finally, the index proposing rate (which reflects the proportion of words proposing a certain index) can be used to assist in prediction of correctness of proposed indexes. It is found that when average index proposing rate of proposed indexes is larger, the precision of these indexes will be increased. The average index proposing rate can reflect the correctness of indexes.

Chapter five

Real case study in database of Chinese Medicinal Material Research Center

In the previous chapter, the performance of automatic indexing procedures have been demonstrated by simulations using imaginary data. Now, a number of real documents selected from the database of the Chinese Medicinal Material Research Center (CMMRC) will be used to perform some procedures of automatic indexing. The aim is to study the result of these procedures applied to the real documents and to look for the reasons for the problems encountered and the solutions to them.

5.1 Selection of real documents

A total of 103 documents have been selected from the CMMRC database. The criteria to choose them is that they are already indexed by the NLM. The documents indexed by the NLM are appropriate to be used as the training documents to suit the circumstance of the CMMRC. In fact, these 103 documents are those which are published by the CMMRC in 1989 (CMMRC published about 1,200 documents that year) and indexed by the NLM from January, 1989 to June, 1990. This selection criteria can prove that there is, indeed, a certain amount of Chinese medical documents indexed by the NLM

so that they can be practically used in the implementation of automatic indexing in the CMMRC.

In these real case studies, words extracted from the titles and the abstracts of the documents will be used in the automatic indexing procedures as those used in the simulations. Since these documents are translated from Chinese, the English version translated by CMMRC may not be identical with that translated by the NLM. In these case studies, version translated by the CMMRC will be used.

In these case studies, this set of documents will be used to perform automatic indexing under some different conditions and assumptions. The results of these studies will be presented in the following paragraphs.

5.2 Case study one: Overall performance using real data

In this case study, each single word will be treated as an independent item. For example, a term "red blood cell" will be spitted into three items "red", "blood" and "cell". They are managed as if they are not related to each other.

In this case study, associations between indexes and words (found in the titles and abstracts) will be calculated and then the documents will be treated as non-indexed documents. The words found in each document will be used to propose indexes. These procedures are as those used in the simulations. The performance of the automatic indexing can be measured by comparing the original indexes with the proposed indexes. The processes of automatic indexing in this case study will be mentioned as follows.

Stopword elimination

After selection of these documents, words found in the stopword list (shown in the Appendix A) will be eliminated first. After stopword elimination, there is no further process controlling the occurrences and forms of words. No word standardization has been implemented in this case study.

Calculation of associations

The associations between indexes and words found in these 103 documents are then calculated. In these documents, there are 384 different indexes and 2,252 different words used to calculate these associations. The procedure of discarding false associations has not been done in this case study because there are only about 100 training documents. The process of discarding false associations can only be implemented if there are enough documents to perform several training phases. Thus, all associations are retained as if they are correct.

Proposing indexes

After the calculation of associations, indexes will be proposed for each document based on the frequencies of words found in it and the calculated associations between words and indexes.

Since the number of training documents is relatively small compared with large number of words and indexes found in them, the statistical errors caused by low occurrence frequencies of words and indexes will be significant. As mentioned before, there are 2,252 different words found in these 103 documents. However, among these words, only 1,314 words appearing two times or more in all documents. If a word appears only one time in all

documents, its $P(i_j/w_i)$ value will be one (ie. the highest value). But this high value may be just due to the statistical fault caused by the low occurrence of the word. For this reason, the index proposing method in this case study is modified as follows. Assume there are n documents in a document set $\{d_1, d_2, d_3, \ldots d_{n-1}, d_n\}$. For a certain document d_i to be indexed, only other n-1 documents $\{d_1, d_2, d_3, \ldots d_{i-1}, d_{i+1}, \ldots d_{n-1}, d_n\}$ will be treated as training documents to establish associations between words and indexes. The indexes proposed for the document d_i will be based on these calculated associations. Therefore, the index-word associations used by each document to propose indexes will not be identical.

Performance measurement

Basically, precision of proposed indexes and recall of original indexes will be used to calculate the performance. The precision (P_{iN}) and recall (R_{iN}) for each document will be defined as follows.

$$P_{iN} = \frac{\text{no. of correct indexes found in first ixN proposed indexes with highest proposing frequencies}}{i \times N} \qquad (13)$$

$$R_{iN} = \frac{\text{no. of correct indexes found in first ixN proposed indexes with highest proposing frequencies}}{N} \qquad (14)$$

where N = no. of original indexes assigned to a certain document.

The values of P_{1N}, $P_{1.5N}$, P_{2N}, R_{1N}, $R_{1.5N}$ and R_{2N} will be calculated for each document. The average values of all 103 documents will be used to reflect the overall performance.

Before the presentation of the result, the estimated performance will be calculated in order to make a comparison between estimated performance and actual one.

As mentioned before, there are 384 different indexes and 2,252 different words found in these 103 documents. All these 384 index terms used in the case studies are shown in the Appendix B. Occurrence frequencies of these words and indexes are different. Some statistical data about occurrence frequencies of these words and indexes are listed below.

No. of different words = 2,252

Sum of occurrence frequencies of all words = 5,484

No. of words occurring more than one time = 938

Sum of occurrence frequencies of these 938 words = 4,170

Proportion of words possible to propose indexes

= 4,170/5,484 = 76.04%

No. of different indexes = 384

Sum of occurrence frequencies of all indexes = 675

No. of indexes occurring more than one time = 91

Sum of occurrence frequencies of these 91 indexes = 382

Proportion of indexes possible to be proposed by words

= 382/675 = 56.59%

On the average, the estimated P_{1N} and R_{1N} values will be about 43.03% (ie. 76.04% x 56.59%). According to this estimation, in this case study, the automatic indexing method can propose 43.03% of correct indexes for each document averagely. The actual result will be normalized according to this estimated result.

Result and analysis

The average values of P_{1N}, $P_{1.5N}$, P_{2N}, R_{1N}, $R_{1.5N}$ and R_{2N} of all 103 documents are presented as follows.

Average precision (after normalization)

Average P_{1N}	Average $P_{1.5N}$	Average P_{2N}
32.68%/43.03% = 75.95%	24.84%/43.03% = 57.73%	20.53%/43.03% = 47.71%

Average recall (after normalization)

Average R_{1N}	Average $R_{1.5N}$	Average R_{2N}
32.68%/43.03% = 75.95%	37.37%/43.03% = 86.85%	41.08%/43.03% = 95.47%

From the above result, it is found that although there is rather small number of training documents used in this case study, the automatic indexing procedure can attain a certain level of performance. On the average, about 76% of indexes possible to be proposed can be found out.

As mentioned before, the estimated performance is about 43%. Of course, this poor performance is mainly caused by the insufficient number of documents which cannot provide enough information to establish the accurate associations between words and indexes. Refer back to the result in the

training phase simulation when the word diversity is uncontrolled, there are only 50 words and 50 indexes allowed to occur in the documents but 1,000 documents are required to attain a very high performance. In this case, there are only 103 documents but many words and indexes found in them. Therefore, the obvious method to improve the performance is to increase the size of training document set.

5.2.1 Sample results of automatic indexing for real documents

Three sample documents selected from 103 real documents of CMMRC and their indexing results by automatic indexing are illustrated. The proposed indexes are determined by the procedures used in case study one. This means that the indexes are proposed by single words which are treated as independent items. The indexing results of these three documents vary from high performance to low performance. The reasons for difference in the indexing performance among these three documents will be explained.

These three sample documents are found in the "Abstract of Chinese Medicines, Vol.3 No.2 1989" published by the CMMRC. Their document numbers in this abstract are listed as below.

Sample A: 890341
Sample B: 890402
Sample C: 890414

As mentioned before, the words used to perform automatic indexing are from the titles and abstracts while the index terms are those assigned by the National Library of Medicine (NLM).

The original index terms, title, abstract and automatic indexing result (ie. proposed indexes and indexing performance) of each sample document are illustrated on next few pages. Then, the explanation for difference in performances among these three documents will be covered.

Indexing performance of sample A

Index terms assigned by NLM
Alkaloids--Isolation and Purification--IP;
Drugs, Chinese Herbal--Analysis--AN;
Glucosides--Isolation and Purification--IP;
Glycosides--Isolation and Purification--IP;
Chemistry

Title
STRUCTURES OF 2 NEW ALKALOIDAL GLUCOSIDES OF NAUCLEA OFFICINALIS.

Abstract
Two new alkaloidal glucosides were isolated and identified from the stem of Nauclea officinalis <Danmu > (Rubiaceae). They were structurally determined by chemical and spectral methods and named as nauclecoside and nauclecosidine. The known alkaloid vincoside lactam was also isolated.

First 6 proposed index terms	P.F.	T.N.
1) Drugs, Chinese Herbal--Analysis--AN ✓	12.40	40
2) Chemistry ✓	10.72	26
3) Mice	03.35	22
4) Alkaloids--Isolation and Purification--IP ✓	03.32	04
5) Glucosides--Isolation and Purification--IP ✓	02.70	02
6) Glycosides--Isolation and Purification--IP ✓	02.70	02

Performance = 80%

Note:

1) Proposed index terms marked with ✓ are correct ones

2) P.F. = proposing frequencies of an index term

3) T.N. = number of training documents indexed by a certain index term = number of documents (found in total 103 documents) indexed by this index term - 1

4) Performance is measured by

$$\frac{\text{no. of correct indexes found in first N proposed indexes}}{N}$$

where N = number of original indexes assigned by NLM

Indexing performance of sample B

Index terms assigned by NLM
Anti-Inflammatory Agents, Non-Steroidal;
Drugs, Chinese Herbal--Pharmacology--PD;
Alcohol, Ethyl;
Mice;
Prostaglandins E--Metabolism--ME;
Rats

Title
STUDIES ON THE ANALGESIC AND ANTI-INFLAMMATORY ACTIONS OF ALTHAEA ROSEA.

Abstract
The 60%-ethanol extract of the corolla of Althaea rosea <Shukuihua > (Malvaceae) was prepared as an aqueous suspension. At 10 g/kg PO in mice, the suspension increased the thresholds of pain caused by acetic acid and radiation heat. The same treatment also decreased acetic acid-induced increase in capillary permeability and in rats dextran-induced paw edema and carrageenin-induced increase in PGE content in paw exudate. A dose of 80 g/kg PO in mice fasted for 12 h reduced their spontaneous activity but did not cause any death in 72 h. Its LD50 was 2.76±0.08 g/kg IV in mice.

First 6 proposed index terms	P.F.	T.N.
1) Mice ✓	32.97	22
2) Rats ✓	26.32	24
3) Drugs, Chinese Herbal--Therapeutic Use--TU	16.78	10
4) Drugs, Chinese Herbal--Pharmacology--PD ✓	16.09	11
5) Inflammation--Drug Therapy--DT	13.88	02
6) Anti-Inflammatory Agents, Non-Steroidal ✓	12.18	04

Missed correct index terms	P.F.	T.N.
Alcohol, Ethyl	-	00
Prostaglandins E--Metabolism--ME	00.17	01

Performance = 67%

Indexing performance of sample C

Index terms assigned by NLM
Alkaloids--Therapeutic Use--TU;
Anti-Inflammatory Agents, Non-Steroidal;
Arthus Phenomenon--Drug Therapy--DT;
Inflammation--Drug Therapy--DT;
Arthritis, Adjuvant--Drug Therapy--DT;
Cell Migration Inhibition;
Hypersensitivity, Delayed--Drug Therapy--DT;
Mice;
Rats

Title
ANTI-INFLAMMATORY AND ANTI-ALLERGIC ACTIONS OF ALOPERINE.

Abstract
Aloperine, an alkaloid of Sophora alopecuroides <Kudouzi >, markedly suppressed rat paw swelling induced by carrageenin, mycostatin, PGE2, histamine, 5-HT and scald. It inhibited leukotaxis and the increase in capillary permeability caused by histamine. Its inhibitory effect on carrageenin-induced rat paw swelling was not abolished by adrenalectomy. It reduced the content of PGE and histamine in the exudate formed after injecting carrageenin and dextran in rats, stabilized erythrocyte membranes, and in mouse intoxicated by ethanol increased the activity of catalase but reduced the content of malondialdehyde in hepatic tissue. It had no apparent effect on the serum activity of superoxide dismutase and phagocytosis of the monocyte-macrophage system in mice, Forssman cutaneous vasculitis and the content of immune complex in serum of rats with Arthus reaction. However, it inhibited PCA reaction, Arthus reaction, reversible passive Arthus reaction, delayed hypersensitivity reaction (induced by tuberculin in rats), and adjuvant arthritis.

First 6 proposed index terms	P.F.	T.N.
1) Mice ✓	36.23	22
2) Rats ✓	33.88	24
3) Drugs, Chinese Herbal--Pharmacology--PD	21.05	11
4) Anti-Inflammatory Agents, Non-Steroidal ✓	15.88	04
5) Rats, Inbred Strains	14.56	11
6) Alkaloids--Pharmacology--PD	11.05	07

Missed correct index terms	P.F.	T.N.
Alkaloids--Therapeutic Use--TU	01.62	03
Arthus Phenomenon--Drug Therapy--DT	-	00
Inflammation--Drug Therapy--DT	05.32	02
Arthritis, Adjuvant--Drug Therapy--DT	-	00
Cell Migration Inhibition	07.38	02
Hypersensitivity, Delayed--Drug Therapy--DT	-	00

Performance = 33%

Explanation of the indexing result of sample documents

For each of these three sample documents, the first few proposed indexes with the highest proposing frequencies have been shown. No matter these proposed indexes are correct or not, it is found that on the average, they are relatively common in many training documents (ie. T.N. higher). Generally speaking, when the proposing frequency of a proposed index is higher, the number of training documents indexed by this index will also be larger. This reflects the feature that when an index is common in the training documents, it can be easily proposed. This index can be associated with more different words found in different documents so that it has higher chance to be proposed.

Conversely, the missed correct indexes are often those uncommon in the training documents (T.N. lower). Their rarities cause them to have lower chances to be associated with more different words to establish proper correlations with words. Thus, they are relatively difficult to be proposed correctly. But in these three sample documents, there are some exceptions which will be explained later.

For the sample A, all correct indexes can be found in first six proposed indexes. Although the 4th to 6th proposed indexes are relatively uncommon in the training documents, they can still be proposed correctly. The main reason is that there is one training document whose indexes and content are very close

to those of the sample A. This training document (say sample AA) is listed as follows. The training document of sample AA is found in the "Abstract of Chinese Medicines, Vol.3 No.3 1989" and its document number is 890695.

Sample AA similar to sample A

Index terms assigned by NLM
Alkaloids--Isolation and Purification--IP;
Drugs, Chinese Herbal--Analysis--AN;
Glucosides--Isolation and Purification--IP;
Glycosides--Isolation and Purification--IP;
Chemistry

Title
STRUCTURE OF PINGBEIDINOSIDE FROM THE STEM AND LEAF OF FRITILLARIA USSURIENSIS.

Abstract
A new steroidal alkaloidal glucoside named pingbeidinoside was isolated from the stem and leaf of Fritillaria ussuriensis <Pingbeimu > (Liliaceae). It was elucidated by chemical and spectroscopic methods as 3ß,16α,20-trihydroxy-Δ5-22,26-epiminocholestane-25-O-ß-D-glucoside.

First, it is found that the indexes of the samples A and AA are exactly the same. Also, there are some important words such as "alkaloidal", "glucoside", and "isolated" commonly found in these two abstracts. The high similarity between a training document and a non-indexed document can lead to a high indexing performance. Because when two documents' contents are close to each other, their indexes will also be similar.

For the sample B, although there are two indexes which are incorrect among the first six proposed indexes, the meanings of these two incorrect indexes are, in fact, very close to those of four correct indexes found in first six proposed indexes. The incorrect index "Drugs, Chinese Herbal--Therapeutic Use--TU" is close to correct index "Drugs, Chinese Herbal--Pharmacology--PD". The incorrect index "Inflammation--Drug Therapy--DT" and correct index "Anti-Inflammatory Agents, Non-Steroidal" deal with something about the concept of "inflammation". The reason for two missed correct indexes, as mentioned before, is their rarities causing them with lower chance to be proposed. The reason for proposing the last two (ie. 5th and 6th) proposed indexes despite their low T.N. values is that the $P(i_j/w_i)$ values between indexes about "inflammation" and words about "inflammation" are rather high. This is because the occurrences of indexes about "inflammation" and those words about "inflammations" are consistent in these 103 documents. On the next page, there are titles of documents indexed by the indexes about "inflammation". The consistency between proper indexes and words can establish correct correlations between them regardless of the occurrence frequencies of them.

Following are titles of documents (found in 103 CMMRC documents) indexed by Anti-Inflammatory Agents, Non-Steroidal

1) STUDIES ON THE ANALGESIC AND ANTI-INFLAMMATORY ACTIONS OF ALTHAEA ROSEA.

2) ANTI-INFLAMMATORY AND ANTI-ALLERGIC ACTIONS OF ALOPERINE.

3) EFFECTS OF TETRANDRINE ON VASCULAR PERMEABILITY AND NEUTROPHIL FUNCTION IN ACUTE INFLAMMATION.

4) PHARMACOLOGICAL STUDIES ON CURCULIGO ORCHIOIDES.

5) ANTI-INFLAMMATORY AND IMMUNOSTIMULATORY ACTIONS OF S-4001.

Following are titles of documents (found in 103 CMMRC documents) indexed by Inflammation--Drug Therapy--DT

1) ANTI-INFLAMMATORY AND ANTI-ALLERGIC ACTIONS OF ALOPERINE.

2) EFFECTS OF TETRANDRINE ON VASCULAR PERMEABILITY AND NEUTROPHIL FUNCTION IN ACUTE INFLAMMATION.

3) EXPERIMENTAL STUDIES ON YIGUAN DECOCTION.

The explanation for the indexing performance of the sample C is similar to that of the sample B. Among the first six proposed indexes, the incorrect indexes have close relationships with correct ones. For example, the incorrect index "Rats, Inbred Strains" is close to two correct indexes "Rats" and "Mice". The incorrect index "Alkaloids--Pharmacology--PD" and the correct index "Alkaloids--Therapeutic Use--TU" mention something about alkaloids. The

reason for the sample C having many missed correct indexes is that all these missed indexes are very uncommon in the training documents. In fact, half the missed indexes cannot be found in the training documents and, thus, cannot be proposed eventually.

Before the end of the discussion about the indexing performance of real documents, the effect of occurrence frequencies of words found in the documents will be mentioned. Like indexes, if a certain word is not common in the training documents, this word will affect the indexing performance. For example, in the sample C, the multi-word term "Arthus reaction" found in the last few lines of the abstract is obviously related to the index "Arthus Phenomenon--Drug Therapy--DT". But this term is only found in this document. In other words, for this document, no training document contains this important item that can have significant correlation with the index. Thus, the rarity of a word can lead to the word having lower chance to be associated with proper indexes.

As mentioned before, the major reason of the low indexing performance, which has been illustrated in these few sample documents, is that the number of the training documents is not enough to cover sufficient words and indexes. Therefore, correct correlations between indexes and words cannot be established. The obvious solution to this problem is to increase the size of the training document set.

5.3 Case study two: Using multi-word terms

In this case study, a meaningful multi-word term will be treated as a complete item. For example, "red blood cell" will be managed as one "word" in the automatic indexing. The aim of this case study is to examine the performance difference caused by different definitions of words. In the case study one, a single word will be treated as an independent item. The results of these two case studies can be compared in order to observe if there is any important difference in the performance. The procedures used in this case study are identical with those used in the case study one.

Before the presentation of the result, the estimated performance will be calculated in advance. When multi-word terms are used as "words", there are 1,570 different words found in all documents. Some statistical data about occurrence frequencies of these words are listed below.

Sum of occurrence frequencies of all words = 2,671

No. of words occurring more than one time = 481

Sum of occurrence frequencies of these 481 words = 1,582

Proportion of words possible to propose indexes

= 1,582 / 2,671 = 59.23%

If one uses the same rationale (used in the case study one) to estimate the average P_{1N} and R_{1N}, they will be about 33.52% (ie. 59.23% x 56.59%).

Result and analysis

The average values of P_{1N}, R_{1N} of all 103 documents calculated in this case study will be compared with those calculated in the case study one in order to examine if there is any difference in the performance caused by different definition of a word.

Average precision (after normalization)

	Average P_{1N}
case study one: using single word	32.68%/43.03% = 75.95%
case study two: using multi-word term	32.17%/33.52% = 95.97%

Average recall (after normalization)

	Average R_{1N}
case study one: using single word	32.68%/43.03% = 75.95%
case study two: using multi-word term	32.17%/33.52% = 95.97%

It is found that the performance of using multi-word terms is obviously better than that of using single words. The reason is that when one uses multi-word terms, the word diversity of a document will be lower. As mentioned before, if the word diversities of documents are low, the performance of

automatic indexing will be improved. For example, in a certain document, after a multi-word term, say "action potential", is spitted into two single words "action" and "potential", the word diversity of this document will be increased. The meanings carried by these two single words will be distinct. Also, the indexes associated with these two single words will be very different since these two words are common components of many multi-word terms.

Nevertheless, the use of multi-word terms has its drawback that the proportion of words possible to propose indexes is reduced from 76.04% (in case study one) to 59.23% (in case study two). The reason is that on the average, the occurrence frequency of a multi-word term will be smaller than that of a single word in the same set of documents.

In this case study, it is found that the use of multi-word terms in the automatic indexing has two opposite effects on the performance. It can lower the word diversities of documents. But it also reduces the proportion of words that can be used to propose indexes. But if the training document number is larger, the drawback of using multi-word terms will be reduced. It is because when the document number increases, the chance for a certain multi-word term to occur more frequently will become higher. Then, the proportion of multi-word terms able to propose index will also become higher.

5.4 Case study three: Using concept headings

In the case study one, there are 2,252 words used. In this case study, a smaller set of concept headings will be attempted. The aim of this case study is to verify the usage of concept headings in real documents. Concept headings will be used to represent each word found in these documents and will be used to perform the automatic indexing.

Determination of concept headings

Before using concept headings to perform any process, each word should be determined to be represented by which concept headings. In this case study, the criteria to assign concept headings to words is that if an index, say index X, has occurred in several documents, words of these documents will share a certain concept heading that denotes connection with an index X. Therefore, if a word is very common in many documents indexed by different indexes, this word will be represented by numerous concept headings. The purpose of this criteria is to confirm that all or some of concept headings representing each word will be shared by other words in different documents. (Assume there is no document whose indexes are entirely unique for it.) In this case study, the number of concept headings used is fewer than 400.

The procedures used in this case study are similar to those in the case study one. But in this case study, words will be converted into corresponding concept headings to perform the training and indexing phases. In the training phase, the associations between indexes and concept headings (rather than words) will be calculated. In the indexing phase, the indexes will be proposed based on the concept headings converted from words found in the documents.

In this case study, only words found in the titles will be used to perform automatic indexing. The automatic indexing will be performed by using words and using concept headings respectively. The difference in the performance can be compared.

Result and analysis

The average values of P_{1N}, $P_{1.5N}$, P_{2N}, R_{1N}, $R_{1.5N}$ and R_{2N} of all 103 documents are presented as follows. The results are not normalized.

Average precision

	Average P_{1N}	Average $P_{1.5N}$	Average P_{2N}
Words	27.43%	19.96%	16.69%
Concept headings	32.98%	25.05%	20.75%

Average recall

	Average R_{1N}	Average $R_{1.5N}$	Average R_{2N}
Words	27.43%	29.97%	33.43%
Concept headings	32.98%	37.66%	41.56%

The result shows that there is a little improvement in the performance after the concept headings are used. Thus, from this result, it is found that a smaller number of concept headings can replace a relatively large number of

words in the automatic indexing. The use of concept headings can solve the problem of managing a large number of natural language terms that appear in the documents. Also, according to this result, it is found that the concept headings can be adapted in the statistical approach of automatic indexing although they are seldom attempted in the past researches.

5.5 Case study four: Prediction of proposed index correctness

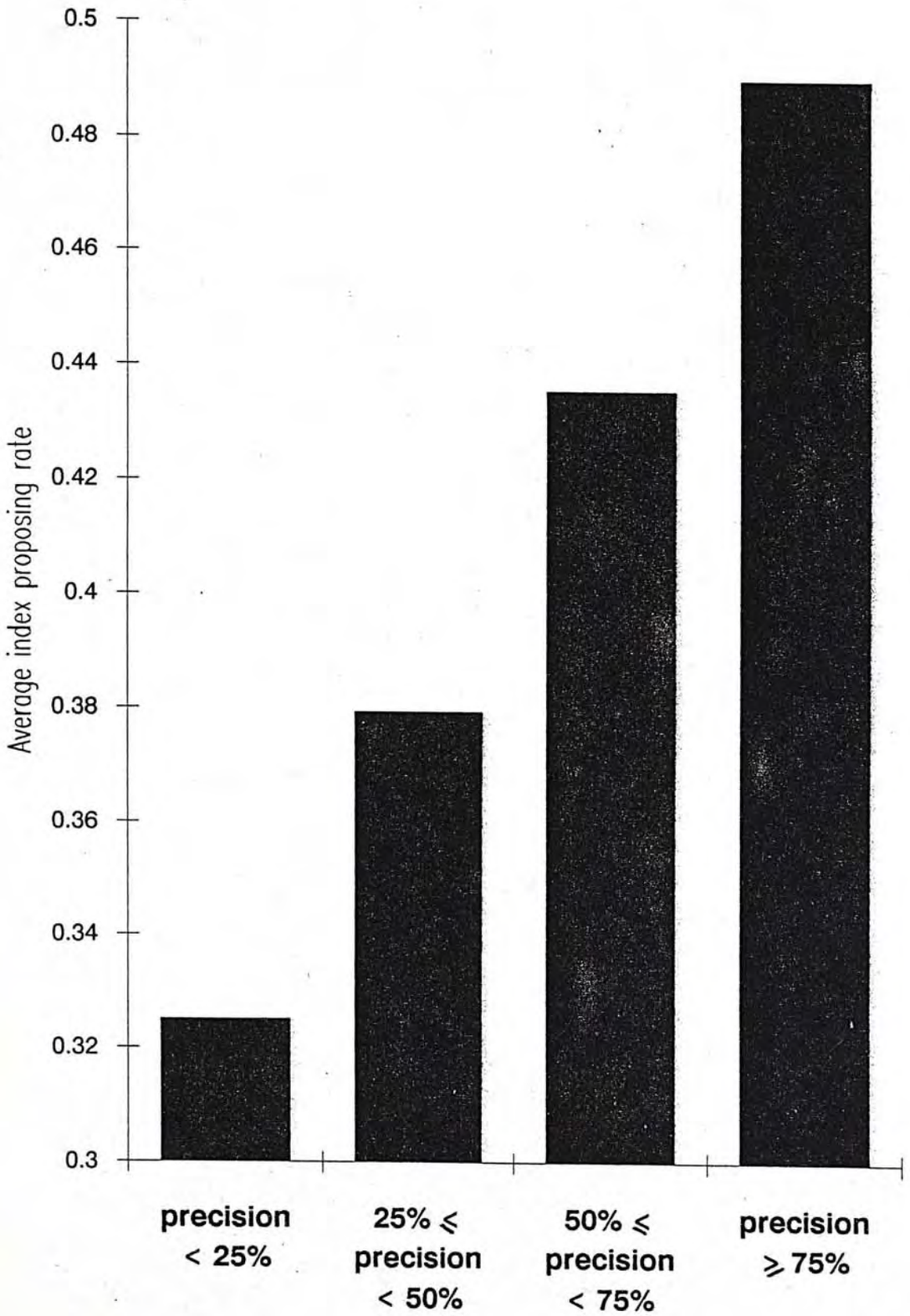
From the results of case study one, it is found that on the average, only about 43.03% of original indexes can be proposed due to insufficient training documents. No matter the proposed indexes are correct or not, the automatic indexing procedures will routinely propose some indexes to each non-indexed document. The index proposing rate of an index is able to predict the correctness of the index. As illustrated in the result of the simulation, there is a correlation between the average index proposing rate of proposed indexes and the precision of them. The aim of this case study is examine whether the index proposing rate can work practically to predict the correctness of proposed indexes in real documents.

The procedures used in the case study one will be repeated but this time the index proposing rate will be calculated for each index proposed for every document.

Result and analysis

The result of this case study is shown on the graph 9 on next page. In this graph, the average index proposing rates of proposed indexes in different ranges of precision are compared. (The precision used in this case study is P_{1N} and is not yet normalized.) It is found that when the precision of proposed

Graph 9
Change of average index proposing rate with different range of precision



indexes get higher, the average index proposing rates of them also become larger. From this result, it is found that the index proposing rate can be used as a hint to reflect the correctness of proposed indexes in real documents.

5.6 Case study five: Use of $(\Sigma\triangle R_{ij})/F_i$ to determine false association

As mentioned before, the process of discarding false associations has not been performed in the case study of real documents of CMMRC. The main reason for not performing this process is that the number of training documents used is rather small compared with a large quantity of words and indexes found in these documents. Many words and indexes occur only one time in these documents so that there is no enough chance for these words to change their ranks according to the change of the $P(i_j/w_i)$ values which are calculated each time when training documents are increased. Without the chance to change the word ranks, it is impossible to use the feature of rank change to determine whether a certain word is truly associated with a certain index.

However, in these documents, some words and indexes having high occurrence frequencies can be used to illustrate that the changes of word ranks can be used to determine whether a word is truly associated with an index. Following are some examples.

Among the 103 training documents, the MeSH term RATS has been used to index 25 documents. In other words, about 1/4 of total training documents have been indexed by RATS. The high occurrence frequency of this index may imply that the words that ought to be associated with this index should have enough chance to change their ranks. Therefore, this index has

been attempted to illustrate the method of discarding false associations (or determining correct associations).

Procedures

All training documents have been divided into ten portions. Each of first nine portions has ten training documents. The last portion has 13 training documents. (The total number of documents found in these ten portions is 103.) There will be ten successive training phases. In each training phase, one portion of training documents will be appended. This means ten documents are used in first training phase, twenty documents in second training phase, and so on. The rank changes of words that are associated with the index RATS and the values of $(\Sigma\Delta R_{ij})/F_i$ of words are calculated as mentioned in the chapter three.

Result and analysis

After ten training phases, the $(\Sigma\Delta R_{ij})/F_i$ of words are calculated. The first two words with the smallest $(\Sigma\Delta R_{ij})/F_i$ are "rat" and "rats". This result is very satisfactory when compared with the result of the traditional method using $P(i_j/w_i)$ to determine the correctness of index-word associations. In the last training phase (ie. all words are used), there are 491 words associated with the index RATS with $P(i_j/w_i)$ value equal to one (the highest statistical correlation

value). Almost all of these 491 words are falsely associated with the index. This large quantity of words having such a high statistical correlation is due to the rarity of them. If a word occurs only one time in all training documents, the statistical correlations between this word and the indexes associated with it is one. However, such a large statistical correlation value is caused by statistical errors due to low occurrence frequencies of words. Therefore, from this example, it is found that the use of $(\Sigma\Delta R_{ij})/F_i$ is better than the use of $P(i_j/w_i)$.

Other example

Similarly, another index ARRHYTHMIA--DRUG THERAPY--DT has been used to evaluate the performance of discarding false associations. This index has been used to index four different training documents respectively. After the last training phase, there are 65 words associated with this index with $P(i_j/w_i)$ value equal to one. However, most of these 65 words are false due to low occurrence frequencies of words. But first two words with the smallest $(\Sigma\Delta R_{ij})/F_i$ value are "sophoramine" and "anti-arrhythmia". The word "sophoramine" is a name of a biochemical used to cure the disease arrhythmia mentioned in some training documents. Also, this example shows that using rank changes of words can assist in the determination of correct associations between indexes and words providing that the words and indexes have, at least, medium occurrence frequencies in the training documents.

5.7 Case study six: Effect of word diversity

As mentioned before, the word diversity of a document can affect the performance of automatic indexing. The results of simulations using imaginary data have illustrated the effect of word diversity on the automatic indexing performance. Now, in the following paragraphs, the effect of word diversity on real documents will be discussed.

In the simulations, the similarity of two words can be controlled and defined deliberately according to the degree of commonness of their predefined indexes or concept headings. The higher the degree of commonness, the higher the word similarity will be. But in the real documents, the occurrences of words found in a document and their associated indexes cannot be controlled. One method used to calculate the word similarity of two words and word diversity of a document is to use the values of $P(i_j/w_i)$ calculated after the training phase. As mentioned earlier, each word associated with n different indexes will be treated as a vector in n-spaces. The cosine correlation between two vectors can be used to reflect the similarity of two words. The word similarity of two words w_a and w_b has been defined (in the chapter three) to be as follows.

$$S_{ab} = \frac{\sum_{j=1}^{n} P(i_j/w_a) \cdot P(i_j/w_b)}{\sqrt{\sum_{j=1}^{n} P(i_j/w_a)^2 \cdot \sum_{j=1}^{n} P(i_j/w_b)^2}} \qquad (4)$$

The average word similarity reflecting the word diversity of a document has been defined (in the chapter three) as follows.

$$AWS = \frac{\sum_{i=1}^{N} \frac{\sum_{j=1, i \neq j}^{N} S_{ij}}{N-1}}{N} \qquad (5)$$

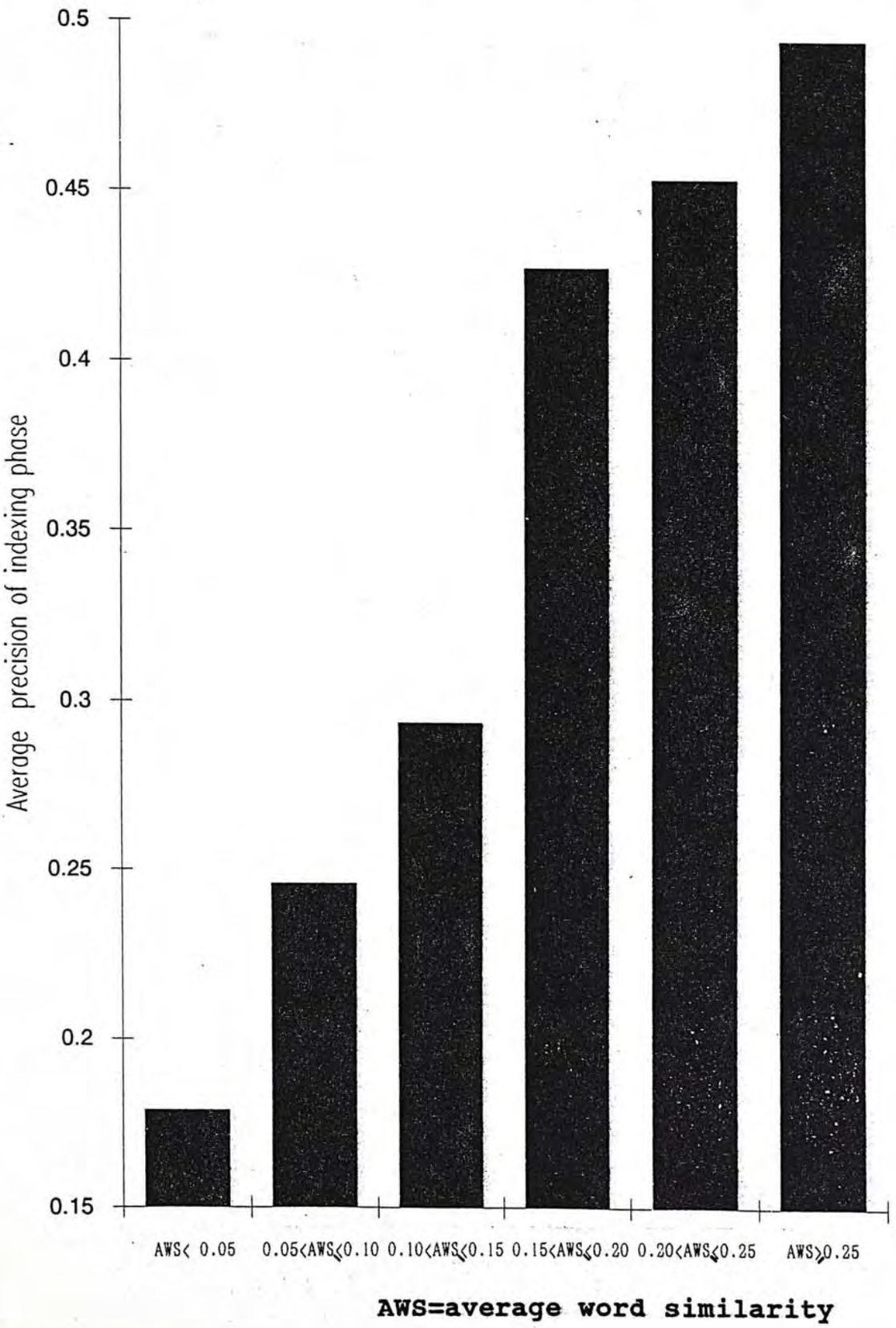
where N is the number of words found in a document.

Result and analysis

The word diversity of training documents used in the case study one have been calculated according to the training phase result. (In this case study, each single word is treated as independent item.) The average word similarity (AWS) of these 103 documents is 0.1331. The relation between the word diversity and the indexing phase performance of the case study one has been shown on the graph 10 on the next page. In this graph, the documents are divided into several groups according to their calculated AWS. The average indexing performance of each group is calculated (the performance is in terms of P_{1N}).

From this result, it is found that the performance of indexing phase is related to the word diversity of documents. The performance is getting better when the word diversities of documents are lower (ie. AWS higher). Thus, this result is consistent with the predicted effect of word diversity and the result of

Graph 10
Change of average precision of indexing phase with word diversity of documents



simulation using imaginary data. When words found in a document are similar to each other, they can concentrate on proposing some common indexes to increase the performance of indexing phase.

5.8 Summary

In these case studies, despite using relatively small amount of documents, many aspects of automatic indexing for real documents in the CMMRC database have been surveyed. First, it is confirmed that there are a certain number of Chinese medical documents indexed by the NLM. They can be adopted as training documents for the implementation of automatic indexing in the CMMRC.

Second, although in these case studies, the processes of word standardization and discarding of false associations have not been performed, the automatic indexing method can suggest a certain quantity of correct indexes that are possible to be proposed under the constraints of these case studies (because some words and indexes occur only one time in all documents).

Third, the use of concept headings is proved to be practical to substitute the words appearing in real documents in the statistical approach of the automatic indexing.

Fourth, it is found that the use of average index proposing rate can assist in the prediction of the precision of proposed indexes. This prediction is important especially when the training document set is small and has not yet covered enough topics, like the situation in these case studies.

Fifth, the use of $(\Sigma\Delta R_{ij})/F_i$ has been verified by some examples. Provided that the frequencies of words and indexes are not too low, the correct index-word associations can be identified by the $(\Sigma\Delta R_{ij})/F_i$ values.

Finally, the word diversities of documents are found to be related to the indexing performance. It is found that the lower word diversity can enhance the performance of suggesting indexes.

Chapter six

Conclusion

In recent years, researchers studying the automatic indexing mainly concentrate on the semantic and syntactic approaches but there are rather few researches on the statistical approach. In this paper, the advantages and significance of the statistical approach are reiterated. Compared with the semantic and syntactic approaches, the statistical approach has by-passed the most difficult problem of creating an indexing algorithm which not only identifies the linguistic entities of natural language, but also understands the meanings conveyed by different combinations of these linguistic entities. In the semantic and syntactic approaches, this intelligent operation is performed by imitating the reasoning process of human being. Typically, the semantic and syntactic rules of natural language will be incorporated in the system to assist in understanding the meanings of natural language. However, incorporating and using these complicated rules in the automatic indexing is not a simple task. It needs expertise which, in turn, means that long development time and high cost are necessary. In the statistical approach, little expertise is required to develop the knowledge component in the system. The knowledge component in the form of statistical correlations between indexes and words can be developed automatically by routine procedures quickly and inexpensively.

Conventionally, in the statistical approach, one difficult problem is to determine which clue words are used to calculate the associations with indexes, and then which calculated associations should be discarded. In the past, there was no objective way to make this decision. Researchers often used the value of statistical correlation between a word and an index to determine whether a certain associations should be deleted. (Large correlation value often means an index-word association is correct.) In fact, using this method has a drawback that the statistical correlation cannot always be a good clue to distinguish correct associations from incorrect ones due to (1) statistical errors caused by low occurrence frequencies of words and indexes, and (2) diverse meanings of words which can link with many different indexes. In this paper, there is a new determination technique using the characteristics that correct words for an index will have lower chance to alter their ranks (determined by comparing the statistical correlation of words that are associated with the index) when the statistical correlations are changed by increasing training documents. This feature can be used to assist in distinguishing correct index-word associations from incorrect ones. The advantage of this method is that it is not completely dependent on the values of statistical correlation which may be erroneous. Simulation results and case study results have shown that this method can assist in determining the boundary between correct associations and incorrect ones clearly.

Semantic representation of natural language is commonly used in the semantic approaches. In this research, the use of semantic representation in the statistical approach has been attempted although it is seldom used in this approach. An advantage of using semantic representation is that a large number of natural language terms can be replaced by a relatively small number of terms of semantic representation. Therefore, this can solve the problem of managing a large number of natural language terms. Moreover, the use of semantic representation can settle the problem of representing synonyms and hierarchial-related words in the automatic indexing. Typically, these two problems are often the critical deficiencies of the statistical approach. In this study, the results of simulations and case studies have shown that it is feasible to use semantic representation in the statistical approach to solve these problems.

According to procedures of the automatic indexing, there is always a certain number of indexes proposed to a non-indexed document routinely, no matter the proposed indexes are correct or not. In this paper, a simple method used to predict the precision of proposed indexes has been suggested. In a non-indexed document, the proportion of words proposing a correct index should be larger than that proposing an incorrect one. An index proposing rate is defined to be the proportion of words proposing a certain index. If the index proposing rate of an index is higher, this index will have higher chance to be a correct one. The advantage of this method is that it mainly uses the features of a non-

indexed document to guess the precision of proposed indexes and this method is comparatively independent of the calculation result of the training phase. The results of the simulation and case study have shown that when the precision of proposed indexes is getting higher, the average index proposing rate of proposed indexes will be increased. Thus, the average index proposing rate can be used as a hint to predict the precision of proposed indexes.

In this paper, the concept of word diversity has been introduced. If the categories of words found in a document are restricted and similar, the word diversity of this document is low. If the categories of words are various and different, the word diversity is high. Although the word diversity is an important factor affecting the performance of the automatic indexing, it is seldom noticed in the past researches. In this study, the results of simulations and case study have shown that when the word diversities of documents are low, the overall performance of automatic indexing can get better.

In this study, the generalized problem of automatic indexing in natural language is investigated. The methods described in this paper can practically assist in the indexing task of the free text-based database. Although in this study, the knowledge domain used to test the feasibility of applying the automatic indexing is medical science, this statistical approach, in fact, can be implemented in the free text-based database system of other domains.

However, the past researches show that scientific papers are more suitable to be indexed by the statistical approach.

Appendix A: List of stopwords

a
about
above
across
after
afterwards
again
against
all
almost
alone
along
already
also
although
always
among
amongst
an
and
another
any
anyhow
anyone
anything
anywhere
are
around
as
at
be
become
becomes
becoming
been
before
beforehand
behind
being
below
beside
besides
between
beyond
both
but
by
can
cannot
co
could
down
during
each
eg
either
else
elsewhere
enough
etc
even
ever
every
everyone
everything
everywhere
except
few
first
for
former
formerly
from
further
had
has
have
he
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
him
himself
his
how
however
i
ie
if
in
inc
indeed
into
is
it
its
itself
last
latter
latterly
least
less
ltd
many
may
me
meanwhile
might
more
moreover
most
mostly
much
must
my
myself
namely
neither
never
nevertheless
next
no
nobody
none
nor
not
nothing
now
nowhere
of
off
often
on
once
one
only
onto
or
other
others
otherwise
our
ours
ourselves
out
over
own
per
perhaps
rather
same
seem
seemed
seeming
seems
several
she
should
since
so
some
somehow
someone
something
sometime
sometimes
somewhere
still
such
than
that
the
their
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
thereupon
these
they
this
those
though
through
throughout
thus
to
together
too
toward
towards
under
until
up
upon
us
very
via
was
we
well
were
what
whatever
when
whence
whenever
whenever
where
whereafter
whereas
whereby
wherein
whereupon
whether
which
while
whither
who
whoever
whole
whom
whose
why
will
with
within
without
would
yet
you
your
yours
yourself
yourselves

Appendix B: Index terms used in case studies

There are 384 different indexes (MeSH terms) found in 103 CMMRC documents used in the case studies. Following are these indexes and the number (figure in the parenthesis) of documents indexed by them in these 103 documents.

(02) 6-Ketoprostaglandin F1 alpha--Metabolism--ME
(01) A-23187--Pharmacology--PD
(01) Abortifacient Agents, Non-Steroidal
(01) Acetaminophen
(01) Acetophenones--Isolation and Purification--IP
(01) Acetylcarnitine--Isolation and Purification--IP
(01) Aconite--Analogs and Derivatives--AA
(01) Aconitine--Analogs and Derivatives--AA
(01) Aconitine--Isolation and Purification--IP
(04) Action Potentials--Drug Effects--DE
(01) Adaptation, Physiological--Drug Effects--DE
(01) Adenine--Isolation and Purification--IP
(01) Adenosine Cyclic Monophosphate--Metabolism--ME
(01) Adenosine Triphosphate--Biosynthesis--BI
(01) Adjuvants, Immunologic
(01) Adult
(03) Aged
(01) Aged, 80 and over
(01) Aging--Drug Effects--DE
(01) Alanine Aminotransferase--Analysis--AN
(01) Alcohol, Ethyl
(01) Aldehydes--Chemical Synthesis--CS
(01) Aldose Reductase--Metabolism--ME
(02) Alkaloids--Analysis--AN
(05) Alkaloids--Isolation and Purification--IP
(01) Alkaloids--Pharmacokinetics--PK
(08) Alkaloids--Pharmacology--PD
(04) Alkaloids--Therapeutic Use--TU
(01) Alloxan
(01) Amides--Isolation and Purification--IP
(01) Ammonium Chloride--Poisoning--PO
(01) Amygdaloid Body--Physiopathology--PP
(01) Analgesics--Isolation and Purification--IP
(01) Anisoles--Isolation and Purification--IP
(01) Anisoles--Pharmacology--PD
(02) Anoxia--Drug Therapy--DT
(01) Anthraquinones--Isolation and Purification--IP
(03) Anti-Arrhythmia Agents

(05) Anti-Inflammatory Agents, Non-Steroidal
(01) Antibiotics, Antineoplastic--Chemical Synthesis--CS
(01) Antibiotics, Antineoplastic--Therapeutic Use--TU
(01) Antibody-Producing Cells--Drug Effects--DE
(01) Anticoagulants--Analysis--AN
(01) Anticonvulsants
(01) Anticonvulsants--Pharmacology--PD
(01) Anticonvulsants--Therapeutic Use--TU
(01) Antimalarials
(01) Antimalarials--Pharmacokinetics--PK
(01) Antimalarials--Pharmacology--PD
(01) Antineoplastic Agents, Phytogenic
(02) Antineoplastic Agents, Phytogenic--Analysis--AN
(02) Antineoplastic Agents, Phytogenic--Isolation and Purification--IP
(01) Antioxidants
(01) Antiviral Agents
(01) Aorta--Cytology--CY
(01) Aorta--Metabolism--ME
(03) Arachidonic Acids--Metabolism--ME
(04) Arrhythmia--Drug Therapy--DT
(01) Arrhythmia--Etiology--ET
(01) Arsenic--Poisoning--PO
(01) Arthritis, Adjuvant--Drug Therapy--DT
(01) Arthus Phenomenon--Drug Therapy--DT
(01) Ascomycetes
(01) Basidiomycetes
(01) Benzaldehydes--Pharmacology--PD
(01) Benzaldehydes--Therapeutic Use--TU
(01) Benzopyrans--Isolation and Purification--IP
(01) Berberine--Analogs and Derivatives--AA
(02) Berberine--Analysis--AN
(01) Berberine--Therapeutic Use--TU
(02) Berbines--Analysis--AN
(01) Berbines--Therapeutic Use--TU
(02) Bicyclo Compounds--Chemical Synthesis--CS
(01) Blood Glucose--Metabolism--ME
(02) Blood Platelets--Metabolism--ME
(02) Blood Pressure--Drug Effects--DE
(01) Blood Viscosity--Drug Effects--DE
(01) Bornanes--Administration and Dosage--AD
(01) Bornanes--Pharmacokinetics--PK
(02) Bridged Compounds--Chemical Synthesis--CS
(01) Bundle of His--Physiology--PH
(02) Calcium Oxalate--Analysis--AN
(01) Calmodulin--Antagonists and Inhibitors--AI
(02) Capillary Permeability--Drug Effects--DE

(01) Capsules
(02) Carbon Tetrachloride
(01) Carbon Tetrachloride Poisoning--Drug Therapy--DT
(01) Carbon Tetrachloride Poisoning--Prevention and Control--PC
(01) Carcinoma 256, Walker--Drug Therapy--DT
(01) Cardiovascular Agents
(01) Cats
(01) Cattle
(01) Cell Count
(03) Cell Migration Inhibition
(01) Cells, Cultured
(01) Cevanes--Isolation and Purification--IP
(01) Charcoal--Pharmacokinetics--PK
(27) Chemistry
(03) Chromatography, High Pressure Liquid
(03) Chromatography, Thin Layer
(01) Chromosome Aberrations--Drug Effects--DE
(01) Chronic Disease
(01) Congo Red--Pharmacokinetics--PK
(01) Constipation--Drug Therapy--DT
(01) Contraceptive Agents, Male--Chemical Synthesis--CS
(03) Coumarins--Isolation and Purification--IP
(02) Coumarins--Pharmacology--PD
(02) Crystallography
(01) Cyclohexanes--Isolation and Purification--IP
(01) Cyclophosphamide--Toxicity--TO
(01) Cytochrome P-450--Metabolism--ME
(01) DNA, Neoplasm--Biosynthesis--BI
(01) DNA, Neoplasm--Drug Effects--DE
(02) Densitometry
(01) Deoxyadenosines--Isolation and Purification--IP
(01) Depression, Chemical
(01) Diabetes Mellitus, Experimental--Chemically Induced--CI
(01) Diabetes Mellitus, Experimental--Drug Therapy--DT
(01) Dimethylnitrosamine--Analogs and Derivatives--AA
(01) Diosgenin--Analysis--AN
(01) Dioxoles--Isolation and Purification--IP
(01) Disease Models, Animal
(01) Diterpenes--Analysis--AN
(04) Diterpenes--Isolation and Purification--IP
(02) Dogs
(04) Dose-Response Relationship, Drug
(02) Drug Combinations
(01) Drug Combinations--Analysis--AN
(02) Drug Contamination
(01) Drug Synergism

(01) Drugs, Chinese Herbal
(41) Drugs, Chinese Herbal--Analysis--AN
(12) Drugs, Chinese Herbal--Pharmacology--PD
(11) Drugs, Chinese Herbal--Therapeutic Use--TU
(01) Electrocardiography
(01) Endothelium, Vascular--Metabolism--ME
(01) Epilepsy--Drug Therapy--DT
(01) Epilepsy--Prevention and Control--PC
(01) Epinephrine--Pharmacology--PD
(01) Ergosterol--Isolation and Purification--IP
(01) Erythrocytes--Drug Effects--DE
(01) Esophageal Neoplasms--Chemically Induced--CI
(01) Esophageal Neoplasms--Prevention and Control--PC
(01) Etoposide--Pharmacology--PD
(01) Fatigue--Drug Therapy--DT
(02) Fatty Acids, Unsaturated--Biosynthesis--BI
(01) Fatty Acids, Unsaturated--Metabolism--ME
(01) Fatty Alcohols--Isolation and Purification--IP
(01) Fertility--Drug Effects--DE
(01) Fibrosis
(01) Flavones--Analysis--AN
(06) Flavones--Isolation and Purification--IP
(01) Flavones--Pharmacology--PD
(01) Frangula--Analysis--AN
(01) Free Radicals
(01) Furaldehyde--Analogs and Derivatives--AA
(01) Furaldehyde--Isolation and Purification--IP
(01) Ginseng
(03) Ginseng--Analysis--AN
(01) Glucans--Isolation and Purification--IP
(01) Glucans--Pharmacology--PD
(03) Glucosides--Isolation and Purification--IP
(03) Glycosides--Isolation and Purification--IP
(01) Glycyrrhiza
(02) Glycyrrhiza--Analysis--AN
(01) Glycyrrhiza--Classification--CL
(02) Gossypol--Analogs and Derivatives--AA
(01) Gossypol--Chemical Synthesis--CS
(01) Gossypol--Pharmacokinetics--PK
(02) Gossypol--Pharmacology--PD
(01) Graft vs Host Reaction--Drug Effects--DE
(05) Guinea Pigs
(01) Heart Function Tests
(01) Heart Rate--Drug Effects--DE
(01) Heart--Drug Effects--DE
(01) Heart--Physiology--PH

(01) Heat
(01) Hemodynamics
(01) Hemodynamics--Drug Effects--DE
(01) Hemolysins--Biosynthesis--BI
(01) Hemolysis--Drug Effects--DE
(01) Hemorrhagic Fever Virus, Epidemic--Drug Effects--DE
(01) Hepatitis, Toxic--Drug Therapy--DT
(03) Hepatitis, Toxic--Etiology--ET
(01) Hepatitis, Toxic--Pathology--PA
(02) Hepatitis, Toxic--Prevention and Control--PC
(01) Hepatitis, Toxic--Therapy--TH
(02) Hydroxyeicosatetraenoic Acids--Biosynthesis--BI
(01) Hydroxyeicosatetraenoic Acids--Metabolism--ME
(01) Hyperlipidemia--Blood--BL
(01) Hyperlipidemia--Drug Therapy--DT
(01) Hypersensitivity, Delayed
(01) Hypersensitivity, Delayed--Drug Therapy--DT
(01) Hypoglycemic Agents
(01) Hypotension--Chemically Induced--CI
(01) Hypotension--Drug Therapy--DT
(01) IgG--Biosynthesis--BI
(01) IgM--Biosynthesis--BI
(01) Indoles--Chemical Synthesis--CS
(01) Indoles--Therapeutic Use--TU
(03) Inflammation--Drug Therapy--DT
(01) Injections, Intraperitoneal
(01) Insomnia--Drug Therapy--DT
(02) Isoflavones--Pharmacology--PD
(02) Isoquinolines--Pharmacology--PD
(01) Kindling (Neurology)--Drug Effects--DE
(02) Lactones--Isolation and Purification--IP
(01) Lactones--Pharmacology--PD
(03) Legumes
(01) Legumes--Analysis--AN
(01) Legumes--Ultrastructure--UL
(01) Lens, Crystalline--Enzymology--EN
(01) Lepidoptera
(02) Lethal Dose 50
(01) Leukemia, Experimental--Drug Therapy--DT
(01) Leukemia, Experimental--Metabolism--ME
(01) Leukocytes--Drug Effects--DE
(01) Leukotrienes B--Biosynthesis--BI
(01) Leukotrienes B--Metabolism--ME
(01) Leydig Cells--Drug Effects--DE
(01) Lignin--Analysis--AN
(01) Lignin--Isolation and Purification--IP

(03) Lipid Peroxidation--Drug Effects--DE
(01) Liver Function Tests
(01) Liver Regeneration--Drug Effects--DE
(01) Liver--Cytology--CY
(01) Liver--Pathology--PA
(02) Macrophages--Drug Effects--DE
(01) Macrophages--Metabolism--ME
(01) Malaria--Metabolism--ME
(01) Malonates--Isolation and Purification--IP
(01) Malondialdehyde--Blood--BL
(01) Malondialdehyde--Metabolism--ME
(02) Mass Fragmentography
(01) Materia Medica
(03) Medicine, Chinese Traditional
(01) Membrane Potentials--Drug Effects--DE
(01) Mesenteric Arteries--Drug Effects--DE
(02) Metabolic Clearance Rate--Drug Effects--DE
(01) Methods
(23) Mice
(01) Mice Mice, Inbred C57BL Necrosis
(02) Mice, Inbred C57BL
(01) Mice, Inbred ICR
(01) Microcirculation--Drug Effects--DE
(01) Microscopy, Electron, Scanning
(01) Microsomes, Liver--Drug Effects--DE
(03) Middle Age
(01) Minerals--Toxicity--TO
(02) Miotics--Chemical Synthesis--CS
(01) Mitochondria, Heart--Drug Effects--DE
(03) Molecular Conformation
(01) Muscle Contraction--Drug Effects--DE
(01) Muscle, Smooth--Drug Effects--DE
(03) Myocardial Contraction--Drug Effects--DE
(01) Myocardial Infarction--Complications--CO
(01) Myocardial Reperfusion Injury--Complications--CO
(01) Myocardium--Cytology--CY
(01) Neoplasm Proteins--Biosynthesis--BI
(01) Neoplasm Proteins--Drug Effects--DE
(01) Neoplasm Transplantation
(01) Neutrophils--Drug Effects--DE
(01) Neutrophils--Metabolism--ME
(01) Nuclear Magnetic Resonance--Methods--MT
(04) Oils, Volatile--Analysis--AN
(01) Oils, Volatile--Therapeutic Use--TU
(01) Oleic Acids
(01) Pain--Physiopathology--PP

(01) Papaverine--Pharmacology--PD
(01) Papillary Muscles--Drug Effects--DE
(01) Papillary Muscles--Physiology--PH
(01) Parasympatholytics--Chemical Synthesis--CS
(01) Parasympathomimetics--Chemical Synthesis--CS
(01) Peritoneal Cavity--Cytology--CY
(02) Phagocytosis--Drug Effects--DE
(01) Pharmacognosy
(01) Phenanthrenes--Analysis--AN
(01) Phenanthrenes--Isolation and Purification--IP
(01) Phenanthrolines--Analysis--AN
(01) Phenols--Chemical Synthesis--CS
(01) Phenytoin--Pharmacology--PD
(01) Plant Extracts
(01) Plants, Medicinal--Analysis--AN
(02) Plants, Medicinal--Anatomy and Histology--AH
(01) Plants, Medicinal--Growth and Development--GD
(01) Plants, Medicinal--Ultrastructure--UL
(01) Plants, Toxic--Analysis--AN
(01) Plasmodium Berghei--Drug Effects--DE
(01) Platelet Activating Factor
(01) Platelet Activating Factor--Antagonists and Inhibitors--AI
(02) Platelet Aggregation Inhibitors
(01) Platelet Aggregation Inhibitors--Isolation and Purification--IP
(03) Platelet Aggregation--Drug Effects--DE
(01) Plethysmography, Impedance
(01) Podophyllotoxin--Analogs and Derivatives--AA
(01) Podophyllotoxin--Pharmacology--PD
(01) Polarography--Methods--MT
(01) Pollen
(01) Polycyclic Hydrocarbons--Pharmacology--PD
(01) Polyporaceae
(01) Polysaccharides--Isolation and Purification--IP
(01) Polysaccharides--Pharmacology--PD
(01) Procainamide--Therapeutic Use--TU
(02) Prostaglandins E--Metabolism--ME
(01) Prostaglandins--Metabolism--ME
(01) Pulmonary Edema--Chemically Induced--CI
(01) Pulmonary Edema--Drug Therapy--DT
(01) Pulmonary Heart Disease--Drug Therapy--DT
(01) Purkinje Fibers--Physiology--PH
(01) Pyrazines--Therapeutic Use--TU
(01) Quercetin--Analogs and Derivatives--AA
(02) Quercetin--Isolation and Purification--IP
(01) RNA, Neoplasm--Biosynthesis--BI
(01) RNA, Neoplasm--Drug Effects--DE Tumor Cells, Cultured

(06) Rabbits
(25) Rats
(12) Rats, Inbred Strains
(01) Receptors, LH--Drug Effects--DE
(01) Regression Analysis
(01) Respiratory Distress Syndrome, Adult--Chemically Induced--CI
(01) Respiratory Distress Syndrome, Adult--Drug Therapy--DT
(02) Review, Tutorial
(01) Rheology
(01) Rutin--Isolation and Purification--IP
(01) SRS-A--Metabolism--ME
(01) Salicylic Acids--Administration and Dosage--AD
(01) Salicylic Acids--Pharmacokinetics--PK
(01) Sapogenins--Analysis--AN
(01) Saponins--Analysis--AN
(03) Saponins--Isolation and Purification--IP
(01) Saponins--Pharmacology--PD
(01) Sarcoma 180--Drug Therapy--DT
(01) Seeds--Anatomy and Histology--AH
(01) Seeds--Classification--CL
(01) Seeds--Ultrastructure--UL
(01) Sensory Thresholds--Drug Effects--DE
(01) Sesquiterpenes--Analysis--AN
(02) Sesquiterpenes--Isolation and Purification--IP
(01) Sesquiterpenes--Pharmacokinetics--PK
(01) Sesquiterpenes--Pharmacology--PD
(02) Shikimic Acid--Isolation and Purification--IP
(01) Shock, Septic--Complications--CO
(01) Shock, Septic--Drug Therapy--DT
(01) Silymarin--Pharmacology--PD
(01) Sinoatrial Node--Cytology--CY
(01) Sinoatrial Node--Drug Effects--DE
(04) Sitosterols--Isolation and Purification--IP
(01) Skin Absorption
(03) Solanaceous Alkaloids--Pharmacology--PD
(01) Solanaceous Alkaloids--Therapeutic Use--TU
(01) Solubility
(02) Species Specificity
(01) Spectrum Analysis, Mass
(01) Spermatocidal Agents
(01) Spirostans--Analysis--AN
(01) Spleen--Immunology--IM
(03) Stereoisomers
(01) Stimulation, Chemical
(01) Subcellular Fractions--Metabolism--ME
(01) Succinates--Isolation and Purification--IP

(01) Sugar Alcohol Dehydrogenases--Metabolism--ME
(01) Tachycardia, Supraventricular--Drug Therapy--DT
(01) Tea
(01) Terpenes--Analysis--AN
(01) Testis--Metabolism--ME
(01) Testosterone--Blood--BL
(01) Thioacetamide
(01) Thromboxane A2--Biosynthesis--BI
(01) Thromboxane B2--Biosynthesis--BI
(01) Thromboxane B2--Blood--BL
(01) Thromboxane B2--Metabolism--ME
(01) Thymoma--Immunology--IM
(01) Thymus Neoplasms--Immunology--IM
(01) Time Factors
(02) Tissue Distribution
(02) Trees
(05) Triterpenes--Isolation and Purification--IP
(01) Tritium
(03) Vasodilator Agents
(01) Vasodilator Agents--Therapeutic Use--TU
(02) Ventricular Fibrillation--Drug Therapy--DT
(01) Ventricular Fibrillation--Etiology--ET
(01) Ventricular Fibrillation--Prevention and Control--PC
(01) Weather

References

[1] Booth, A.D. A 'Law' of occurrences for words of low frequency. *Information and Control* 10, (1967), 386-393.

[2] Chang, H.M., Day, J.J. and Lee, W.S. "Chinese Information on medicinal materials computerisation project." In *Proceedings of The Fifteenth Hawaii International Conference on System Sciences* (Honolulu, Hawaii, USA, Jan. 6-Jan. 8, 1982). Elsevier Science Publishers B.V., North-Holland, 1984, pp. 141-149.

[3] Cleveland, D.B. *Introduction to indexing and abstracting.* Libraries Unlimited, Englewood, 1990.

[4] Dillon, M., and Gray, A.S. FASIT: A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science* 34, 2 (1983), 99-108.

[5] Goffman, W. A searching procedure for information retrieval. *Information Storage and Retrieval* 2, (1964), 73-78.

[6] Hamill, K.A., and Zamora, A. The use of titles for automatic document classification. *Journal of the American Society for Information Science* 31, (1980), 396-402.

[7] Humphrey, S.M., and Miller N.E. Knowledge-based indexing of the medical literature: the indexing aid project. *Journal of the American Society for Information Science* 38, 3(1987), 184-196.

[8] Janas, J.M. Automatic recognition of the part-of-speech for English texts. *Information Processing and Management* 13, (1977), 205-213.

[9] Lindberg, D.A.B. (Ed) *Index Medicus* 31, 1, 1990.

[10] Lindberg, D.A.B. (Ed) *Medical Subject Headings* 31, 1990.

[11] Luhn, H.P. A statistical approach to mechanized encoding and searching of literary Information. *IBM Journal of Research and Development* 1, (1957), 309-317.

[12] Maeda, T. An automatic method for extracting significant phrases in scientific or technical documents. *Information Processing and Management* 16, 3 (1980), 119-127.

[13] Maron. M.E. Automatic indexing: an experimental inquiry. *Journal of the Association for Computing Machinery* 8, (1961), 404-417.

[14] Ming, O. *Chinese-English Manual of Common-Used in Traditional Chinese Medicine*. Guangdong Science and Technology Publishing House, China, 1989.

[15] Pao, M.L. Automatic text analysis based on transition phenomena of word occurrences. *Journal of the American Society for Information Science* 29, (1978), 121-24.

[16] Rowbottom M.E., and Willett P. The effect of subject matter on the automatic indexing of full text. *Journal of the American Society for Information Science* 33, (1982), 139-141.

[17] Rowley, J.E. *Abstracting and indexing*. Clive Bingley, London, 1988.

[18] Sager. N. Sublanguage grammers in science information processing. *Journal of the American Society for Information Science* 26, (1975), 10-16.

[19] Steinacker, I. Indexing and automatic significance analysis. *Journal of the American Society for Information Science* 25, (1974), 237-241.

[20] Trubkin, L. Auto-indexing of the 1971-77 ABI/INFORM database. *Database* June, (1979), 56-61.

[21] Vickery, B.C. Analysis of information. *Encyclopedia of Library and Information Science* 1, (1968), 355-384.

[22] Vleduts-Stokolov, N. Concept recognition in an automatic text-processing system for the life Sciences. *Journal of the American Society for Information Science* 38, 4 (1987), 269-287.

[23] Vleduts-Stokolov, N. An automatic support to indexing a life sciences data base. *Information Processing and Management* 18, (1982), 313-321.

[24] Zipf, G.K. *Human behaviour and the principle of least effort*. Addison-Wesley, Cambridge, 1949.