The Effects of Size on the Function of an Information Retrieval Document Collection

Thesis submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

Brian G. Mushens

Computing Laboratory, University of Newcastle upon Tyne
May 1982

ABSTRACT

A feature of research into Information Retrieval has been the continued use of small test collections in experiments. The assumption that any results will remain valid when the system is used to interrogate a large operational database is examined critically particularly with regard to the difference in size of collections involved and the reasons for this.

Experiments investigating the effects of size on the MEDLARS database with reference to several sub-collections containing varying numbers of documents are described. These include analyses of single term and two-term combination behaviour and actual retrieval searches. The effect on the clustering structure of different small sub-collections is also studied. The results obtained for MEDLARS are examined in the context of some well-known test collections, namely Cranfield 2 and INSPEC.

Results for MEDLARS data indicate that very large collections (> 20,000 documents) may be necessary in order to ensure that the experimental data is indeed representative and

may therefore be used to accurately predict the performance of a particular system in the operational ervircnment.

ACKNOWLEDGEMENTS

My chief debt is to Elizabeth Barraclough, my supervisor, for her enthusiastic involvement and invaluable guidance throughout the course of this research.

My thanks are due to collegues and staff of the Computing Laboratory, especially Nick Rossiter who patiently made the MEDLARS data available whenever it was required.

My wife, Christine, deserves special credit for her ceaseless understanding and encouraging smile particularly through some of the not-so-pleasant periods of my research.

Throughout the period October 1978 to September 1981 the author was supported by a D.E.S./British Library Research Studentship in Information Science.

NOTES

All the experiments described in this thesis were performed on the NUMAC IBM 370/168 running under MTS. All programs including those driving the GHOST graphics package were written in ALGOL W.

This thesis was prepared using the TEXTFORM text-formatting program package.

N.B. TEXTFORM does not have a facility for producing superfixes. The following convention was adopted - "X'N" should be read as "X raised to the power N".

CONTENTS

1 INFORMATION RETRIEVAL

1.1 INTRODUCTION.

The ever increasing technological complexity of the world has resulted in the generation of vast amounts of literature by industry, commerce, government bodies and academic institutions, a phenomenon which has come to be known as the "Information Explosion". The need for research workers to have rapid access to this documentary data has stimulated work in the field of information retrieval (IR).

Information retrieval has a history of some 25 years, when it was realised that the data handling power of computers could be harnessed in order to alleviate the burden imposed on the researcher. Since those early days, IR has progressed steadily, eager to utilise the advances made in the field of computer science.

1.1.1 The context of IR.

"Information Retrieval" is an unfortunate choice of term to describe the process of finding bibliographic references in response to a request for information, because of the all-embracing nature of "information". However, it is generally accepted by workers in the field that information retrieval is synonymous with reference retrieval, and indeed there are very few occasions when "reference retrieval" cannot be used as a substitute. Lancaster (1968) defines IR as follows: "An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request". In other words, an IR system does not answer a request for information directly, but provides a means to enable the user to find the answer i.e. a reference to a document. (A "document" can be a book, paper, report or any other form of written work.) Some systems provide the user with the actual text of the document (document retrieval), but this is becoming increasingly rare due to the high cost of storing a large collection of documents in full text form on even the least expensive computer storage medium.

1.1.2

Seen in this light, information retrieval excludes fact and data retrieval. Fact retrieval systems or question answering systems, as they are often called, relate more to the fields of artificial intelligence and computational linguistics. Such systems are currently restricted to answering questions relating to a simple "world" consisting, say, of blocks of differing sizes, colours and shapes, based on a number of simple assertions about the "world". (See Winograd (1972) for an example of such a system.) Data retrieval is characterised by data base management systems (see Date (1981) for an introduction to the subject). Stock control systems and automatic airline reservations are examples. In this case, the data is restricted to that which is easily quantified and tabulated. Similarly, the query is completely and unambiguously specified by the use of an artificial language which enables an exact match between query and data to be made. Therefore, there can be only one answer.

1.1.3

IR lies somewhere between these extremes, the freedom of fact retrieval and the constraints of data retrieval.

Of course, the data in IF is restricted to documents, but documents are so varied both in form and especially in content that a restriction hardly exists at all. The queries posed to IR systems are ideally expressed in natural language. This enables the naive user to formulate his query more easily, but other forms of query such as the Boolean combination of key words have been successfully mastered. In IR there is unlikely to be one particular document which will satisfy the user's information need. It is more likely that several documents will prove useful to the user to varying degrees. As a result of this, there is no exact match between query and response, no correct answer. The relationship is more of a partial matching - a document will answer the query to some extent but not completely. That is to say, a document can be seen as relevant to a request. This notion of reference is of central importance in the field of IR and warrants further discussion.

1.2 THE PROBLEM OF IR.

Consider the medical researcher who wishes to find out about the effects of a particular drug on rats. The problem is deceptively simple. All he needs to do is read

the documents covering his subject area, medicine in this case, and select those which are of interest to his query. By doing this, he can reject all those documents not relating to his request and therefore achieve "perfect retrieval" - all the documents he selects are relevant. On the other hand, the number of documents covering the field of medicine runs into millions and the task of reading every one would prove extremely tedious and time-wasting, even if it could be fitted into a lifetime.

It is this very problem that IR wishes to solve. Unfortunately, the computer cannot perform the same intellectual processes as a human being. It cannot read and understand the contents of a document, nor determine whether that document is relevant to a particular request for information or not. All it can do is attempt to approximate to the process. Because of this shortcoming, the computer system often retrieves non-relevant documents as well as relevant documents and indeed it is one of the major goals in IR research to minimise the number of non-relevant documents retrieved and maximise the number of relevant documents retrieved, that is, obtain an even closer approximation to the human process, which is capable of perfect retrieval.

1.3 RETRIEVAL EFFECTIVENESS.

The performance of a system can be measured from this angle and this has led to the introduction of parameters to evaluate retrieval effectiveness. The two most commonly used are recall and precision.

$$\text{Recall} = \frac{\text{the number of relevant documents retrieved}}{\text{the total number of relevant documents}}$$

$$\text{Precision} = \frac{\text{the number of relevant documents retrieved}}{\text{the total number of documents retrieved}}$$

Unfortunately, although not a hard and fast rule of IR (Cleverdon (1972), in general, there exists an inverse relationship between the two. Any attempt to improve recall by retrieving more documents reduces precision by increasing the the number of non-relevant documents retrieved. Similarly, if the number of documents retrieved were to be reduced to improve precision, recall would suffer. To obtain a realistic assessment of a system's performance, the precision is calculated at preset levels of recall (0.1, 0.2 etc.). Improvements to a system would be successful if the precision at these levels of recall were to increase.

Although they are the most popular evaluation measures, recall and precision are not the only ones. Fallout, generality and sensitivity have also been proposed. The ubiquitous contingency table can be used to illustrate their meaning and calculation.

	Relevant	Non-relevant	
Retrieved	a	b	a + b
Not Retrieved	c	d	c + d
	a + c	b + d	

$$\text{Recall} = \frac{a}{a + c}$$

$$\text{Precision} = \frac{a}{a + b}$$

$$\text{Fallout} = \frac{b}{a + c}$$

1.3.1 Efficiency.

The performance of an IR system is not judged solely by its ability to retrieve relevant documents, although this is the most important consideration. The system should also be able to provide the user with documents within a reasonable period of time. This is particularly important in online systems, where the user is conducting his search from a terminal, rather than submitting it for batch processing. The "response time" between the keying in of the search request and the appearance of the first document is often determined by the algorithms used in the system and consequently, the use of overelaborate or complex methods of retrieval, whilst it may lead to an improvement in effectiveness, can adversely affect the response time and increase it to an unacceptable level.

1.4 THE COMPONENTS OF AN IR SYSTEM.

A full explanation of how an IR system works will not be attempted here. A brief overview highlighting in particular the sort of research that has been done in IR, will suffice to give a general impression of its operation. Figure 1.1 shows the basic components of a system diagrammatically.

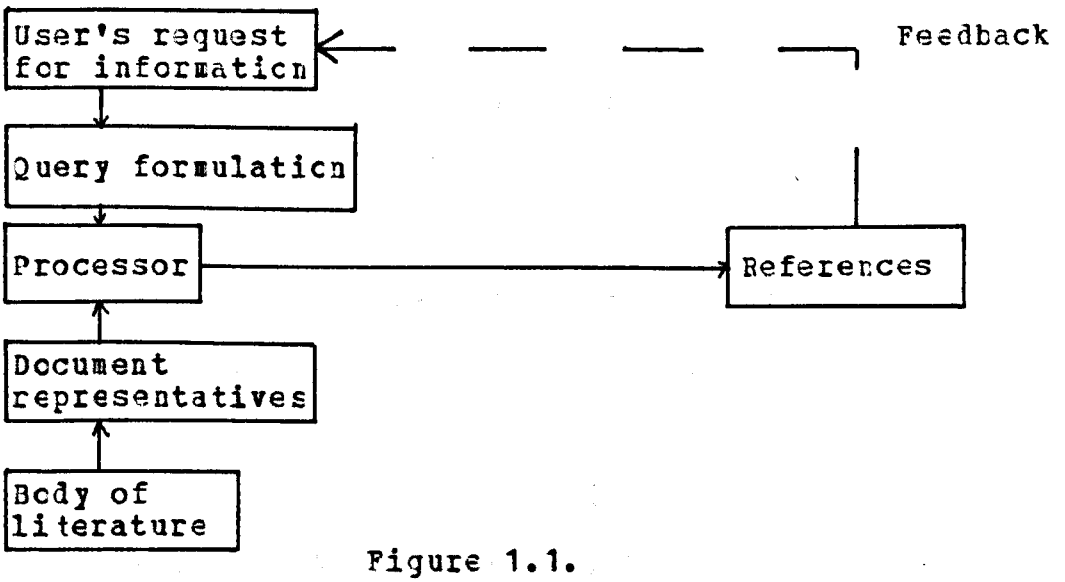


Figure 1.1.

1.4.1 Indexing.

Despite the immense information handling capacity of today's computers, it remains impossible to store the complete texts of even a small number of documents, and even if this problem could be solved the processing of documents in such a form would prove too complex. To overcome these difficulties, the content of each document in a collection is condensed into a form more suitable for computer use.

In a typical IR system, each document is represented by a set of keywords or terms (used synonymously here) according to its content. The assignation of terms to a document can be achieved either manually or automatically.

In manually indexed systems, trained experts assign terms using their knowledge of the subject area. To aid them in this task, a thesaurus of accepted terms is used in order to achieve consistency between indexers. The source for indexing varies from the full text of the document to its abstract or even its title.

The use of experts to index documents is very expensive, because of the nature of the work and the amount of documents that have to be indexed. This has led to much research in the field of automatic indexing. Methods which rely on the semantic and syntactic analysis to extract the content of a document, have proved to be extremely complex in their operation and have shown no significant improvement over relatively simple statistical methods.

Methods which assign terms on a statistical basis use the frequency of terms in the source text (full text, abstract or title). The selection of terms may be made freely from words extracted from the text (free indexing) or controlled by a thesaurus of legitimate terms in much the same way as manual indexing.

An example of an automatic free indexing scheme is given by way of illustration.

1. It has been shown that high frequency words give little indication of the content of a document and therefore as a first step, all non-content bearing words ("fluff" words such as "and", "because", "nevertheless" etc.) are removed from the text.

2. The remaining words are reduced to their stems by suffix stripping. This step is based on the assumption that two words with the same stem refer to the same concept.

3. The resulting set of stems is reduced to the final set of index terms by removing multiple occurrences of the same stem.

The indexing produced by this method is often labelled "binary" indexing, because a term either occurs in a document description or it does not. An elaboration of this method generates document descriptions in which the terms are assigned weights in accordance with their ability to discriminate one document from another and therefore aid effective retrieval.

1.4.2 Search strategies.

There are two main methods of searching collections corresponding to the form of the search request.

Matching function.

Where the user's request for information is expressed in natural language, it is either formulated into a set of search terms by a trained intermediary or automatcally in the same way as a document would be indexed. Once the set of search terms has been obtained, it is compared with the set of document representatives and a ranking of documents is produced according to the degree of association between the document and the query, measured by a matching function (see 4.2.2). This is an attempt by the system to judge the relevance of the documents to the query. The user may then retrieve either the top n documents from the ranking or apply a threshold on the matching function and retrieve all documents exceeding it. By varying either the value of n or the threshold increased recall or precision can be obtained. By lowering the threshold (increasing n), more documents will be retrieved increasing the number of relevant documents retrieved leading to an improvement in recall, but also introducing more non-relevant

documents, which adversely affects precision. By raising the threshold (decreasing n), precision is improved at the expense of recall.

Boolean searching.

In Boolean searching the query is expressed by using logical connectives between terms, e.g.

Q1 = T1 AND (T2 OR T3) AND NOT T4

All documents satisfying the conditions of the search statement are retrieved and are deemed equal in that no ranking is applied to the set of retrieved documents. As such, there is no mechanism available for selective retrieval corresponding to the matching function method.

Boolean searching is efficiently implemented using an inverted file organisation, where each term is associated with a set of documents in which that term occurs. The sets of documents for each term in the search request are processed according to the logical combinations between the terms to produce the set of documents to be retrieved.

1.4.3 Feedback.

In an online environment, the user may submit a query to the system, sample the output and on that basis modify his original query and re-submit it. There have been attempts to automate this process known as relevance feedback.

1.5 OUTLINE OF THE THESIS.

There have been many different methods suggested for the effective retrieval of documents, ranging from the very simplistic to the highly complex. At the present time there are very few occasions where one can say that one particular method is better than the next, because of the lack of conclusive experimental results. This has been due to the wide range of test data sets, which has prevented true comparisons between experiments, and the small number of documents contained in these "test collections" compared to operational IR databases.

Chapter 2 of this thesis studies both types of document collection and highlights the problems concerning the use of "test collections".

Chapter 3 considers the nature of a document collection by theoretical means and by practical experiments studying the effects of size on collection characteristics such as single terms and 2-term combinations.

Chapter 4 analyses the effects of size on the clustering structure of the collection.

Chapter 5 is concerned with the amalgamation of the results of the previous two chapters and how this research affects other work in experimental IR.

Chapter 6 provides suggestions for further research in the area.

2 INFORMATION RETRIEVAL DOCUMENT COLLECTIONS

2.1 INTRODUCTION

An often underestimated component of an IR system is the set of documents upon which that system operates. Indeed, surprisingly little research effort has been directed towards investigating the properties of document collections.

2.1.1

The nature of the documents themselves varies from collection to collection. An average document taken from different collections is indexed by different numbers of terms which, in turn, may have been selected from dictionaries of differing sizes. In addition, the original source for indexing may have been the full text, abstract or title of the document, or perhaps author assigned keywords are used in the document description.

Collections are generally connected with a particular subject area and within the collection, the documents have

all been indexed in a identical manner, using the same source and the same dictionary or thesaurus.

2.1.2

It is common to divide the field of IR into two areas: experimental IR and operational IR. Document collections are no different in this respect. In the operational environment, collections are compiled from the literature pertaining to a particular subject and may be accessed, often along with other databases, by systems offering an information service either to users in-house or to a more widespread population for commercial gain, e.g. DIALOG, BLAISE, INFOLINE, STAIRS.

2.1.3

Experimental IR over the past twenty years has spawned a number of data sets used in the development, testing and evaluation of retrieval methods. These "test collections", as they have become known, have originated in one of two ways. Either they consist of a subset of an operational database, which is made possible by the similarities between documents throughout the database, or

they are specially created by selecting documents from the literature, e.g. Cranfield 2 (Cleverdon et al, 1966).

2.1.4

The assumption governing the use of such test collections is that any experimental results will be reproduced when the system under examination is implemented to access the operational database. In other words, the test data should be representative of the operational data. To determine whether this is true it is necessary to study the characteristics of both types of collection.

2.2 OPERATIONAL DATABASES.

In recent years, there has been a dramatic increase in the number of databases accessible by means of computerised IR systems. There has also been an increase in the number of documents these databases contain and their use in the online environment. Williams (1980) reports that in 1979, there were 528 bibliographic, bibliographic-related and natural language databases, containing some 148 million records. Table 2.1 reproduces

her analysis of U.S. databases.

Size (number of records)	Percentage
< 30K	27
30K-300K	44
> 300K	17
size unknown	12

	100

Table 2.1.

2.2.1

As Table 2.1 shows, the size of a database can range from a few thousand to several million documents. Specific examples are shown in Table 2.2. Ideally, a database should contain sufficient documents to be able to provide the user with an adequate service for that particular subject area. By "adequate service", one would expect that a correctly formatted request for information

Commercial Databases (source: Hall (1977))

Name	Subject Area	Number of Documents
AGRICOLA	Agriculture	1,000,000
CA Condensates	Chemistry	2,260,000
INSPEC	Electrical, electronic, computer & control engineering	350,000
ERIC	Education	250,000
MEDLARS	Medicine	664,000
SCISEARCH	Science	900,000
SOCIAL SCISEARCH	Social Sciences	367,000
TOXLINE	Toxicology	400,000

Table 2.2.

within the scope of the database would relate to at least one document in the collection, however tenuous that relationship may be. As an example, a database claiming to represent the field of medicine would be expected to contain some documents about heart disease.

The number of documents in a database is determined by several factors:

The subject area.

The enormity of the task of constructing a single database to represent all the world's knowledge has

understandably led to the setting up of more practicable databases restricted to more manageable subject areas. The size of database is dependent upon the scope of the subject area - a broad, general subject, such as medicine, in which hundreds of thousands of documents are produced in any one year will require a larger database to represent it than a narrow, more specialised field such as tropical diseases, of which the annual output may be measured in terms of a few thousand documents.

Generally, collections are kept as small as practicable to enable a reasonable response time from the system to be achieved. The compilation is simple in the case of well-defined subjects, such as medicine and electronics. However, where a subject lies in the overlap between two disciplines, the task of deciding whether it is worthwhile combining the two or keeping them separate is more difficult and is usually determined in an ad hoc fashion. Consider the example shown in Figure 2.1.

If one wanted to provide an information service for the field of Pharmacology, it is found that this is partly covered by the field of Medicine and partly by Toxicology. Assuming collections representing

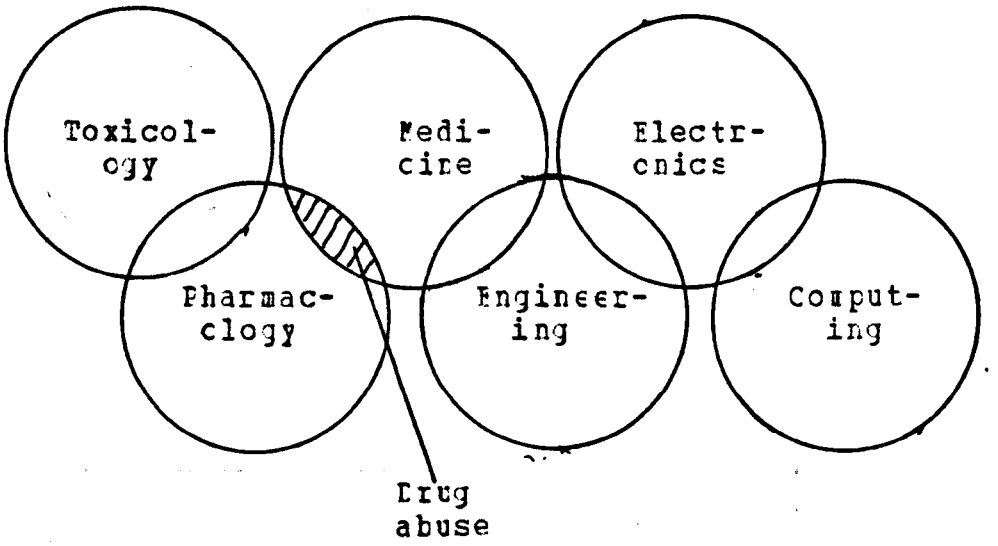


Figure 2.1.

Medicine and Toxicology exist, does one search both databases or one and not the other?

The activity in the area.

The amount of work being performed in a particular field is generally reflected by the number of publications it produces. This can have two contrasting effects on database size. In a thriving subject area, the number of publications will be greater than in one which is in a state of stagnation, thus necessitating, on the face of it, a larger database. On the other hand, if rapid advances are being made in a subject, the literature will become out of date more quickly and need not, therefore, be part of the database. This rate of obsolescence

varies from subject to subject and in their study Burton and Kepler (1960) produced a list of scientific literature "half lives", an analogy with nuclear physics, which is shcwn in Table 2.3.

Literature half lives.

Subject	Years
Chemical Engineering	4.8
Mechanical Engineering	5.2
Metallurgical Engineering	3.9
Mathematics	10.5
Physics	4.6
Chemistry	8.1
Geology	11.8
Physiology	7.2
Botany	10.0

Table 2.3.

Their definition of "half life" is "the time during which one-half of all the currently active literature

was published".

Note that these figures were produced over twenty years ago and due to changes in the nature of the subject areas may not hold true today. Nevertheless, they do give an indication of the variation between subjects. In general, stable sciences exhibit longer half lives than developing sciences, e.g. mathematics vs. physics, whilst applied fields tend to become obsolescent more rapidly than those with a more fundamental and theoretical basis.

Journal selection.

The most common method of compiling a database is to select a number of journals from within the subject area and, upon publication of each issue, include all that issue's papers in the collection. The size of the collection is, in this case, determined by the number of journals included and the duration of coverage. For example, a collection may be made up of the papers contained in a hundred journals over a period of three years. However, this method has the disadvantage of omitting pertinent documents if they appear in journals not directly associated with the subject area, e.g. a medical paper in a general scientific journal such as "Science" may well be

cmitted frcm a medical database.

System function.

The size of a documert collection can also be governed by the purpose for wlich it is to be used. In a retrospective syste, the database should be larg enough to prcvide adequate coverage of the subject (see 2.2.1). A current awareness syste may cnly require a few mcnths' releases of documents and in a purely SDI system only the most up to date release is required.

2.3 IEST_COLIECTIONS

2.3.1

Research in IR over the past twenty years has led to the emergence of a number of sets of data used sclely fcr experimental purpcses, commonly known as "test collections" (Sparck Jcnes & van Rijsbergen (1976)).

In general, a test collection comprises a set of documents, an associated set of queries, and some relevance data. Thus, in a typical experiment each query is submitted to the system, which will in turn retrieve certain documents which it has deemed relevant to that query. The effectiveness of the system can then be judged by comparing the output with the relevance information. In this way it is hoped that the retrieval process in an operational environment may be modelled, the assumption being that any results obtained in the experiment will remain valid.

2.3.2

Present test collections invariably contain a relatively small number of documents, when compared with commercial databases. Table 2.4 gives some examples of test collections and reference to Table 2.2 highlights the difference in size.

There are the following good reasons for restricting the number of documents in test collections, but they must be considered carefully if the validity of the test collection as a representative of the large collection can be called into question.

Test Collections.

Name/ Reference	Subject area	Docs.	Queries
Cranfield 2 Cleverdon et al(1966)	Aeronautics	1400 200	221 42
INSPEC Aitchison et al (1970)	Physics, electrotechnology, control	541	97
ISILT Keen & Digger(1972)	Documentation	797	63
UKCIS Barker et al(1974)	Chemistry	11518 15629	193
MEDUSA Barber et al(1972)	Medicine	51000	58
NPL Vaswani & Cameron (1970)	Electronics, computers, physics, geophysics	11571	93
UKAEA Clive et al(1973)	Nuclear science	12765	60

Table 2.4.

Collection construction.

Compiling a collection can be an expensive and laborious task. The actual gathering of the documents themselves is not difficult (they are plentiful), but they may not be in machine readable form or, even if they are, in a format suitable for the experiment. The construction of the query set poses more of a problem. Robertson (1981) lists four problems concerning query selection. Firstly, queries should be trapped during the short space of time in which the

act of requesting information takes place. Secondly, these acts are performed in a variety of locations, and are therefore difficult to monitor. Thirdly, the subsequent co-operation of the requester is often required, e.g. for relevance assessment, but may not be offered, and fourthly, these difficulties compound to question whether the resulting queries are representative of anything at all.

However, by far the most difficult obstacle to overcome in the construction of a test collection is the provision of relevance information. Current measures of retrieval performance, e.g. recall, precision and fallout (Chapter 1), rely heavily on the knowledge of which documents are relevant to each query. Therefore, to be able to use such measures, each query should be accompanied by a list of documents relevant to it. This can be achieved in two ways. Ideally, the originator of the query should make the assessment as he is best qualified to do so. Unfortunately, this involves a great deal of effort on his part, as many documents will need to be judged. For this reason, experiments often rely on the assessments of subject experts who compare each query with every document and, using their knowledge of the subject, decide on its relevance. However, both these

methods are time consuming and/or expensive an require substantial resources even for small collections.

Ease of experiment.

The use of computers in IR research has meant that experiments can be replicated at will. Even so, computation time and storage requirements remain at a premium and small collections enable experiments to be performed efficiently and quickly, avoiding delays due to re-runs and the incorporation of system modifications. Indeed, it is often possible to have substantial parts of small collections residing in core storage, reducing the overheads of data transfer considerably.

It must be stressed that no matter how convincing the arguments for the use of small test collections are, the overriding requirement for the test collection to be representative of the complete collection should be fulfilled if the results are to be valid. Further, even those researchers who use small test collections express their doubts about doing so, the reasons for which are examined in the next section.

2.4 PROBLEMS CONCERNING THE USE OF TEST COLLECTIONS

2.4.1

In a recent review of IR experiments, Sparck Jones (1981) concludes that twenty years of testing has taught us little about the real nature of retrieval systems and only broad generalisations, mostly unsubstantiated, can be made about the methods of improving their performance.

This lack of conclusive results is blamed upon poor experimental design methodology, to which the unsatisfactory use of test collections can be seen to be a contributory factor.

2.4.2

Particularly during the 1960's and early 1970's, almost every project gave birth to a new test collection, the variety of which was almost as great as the projects themselves. As a consequence of the differences between the test collections, the results of the experiments were for the most part not comparable. This meant that the

recommendations of one experiment were not consolidated by later experiments performed in the same area of inquiry. Examples of the differences between the collections are given below:

1. Subject area - mostly "hard" scientific subjects but wide variation within this description.

2. Indexing source - varying from full text through abstract to title.

3. Indexing method - automatic vs. manual.

4. Indexing language - document derived keywords vs. fully controlled thesauri.

5. Method of creation - the Cranfield 2 documents were specially selected whereas UKCIS used a slice of the current operational collection. Similarly, INSPEC requests were obtained exclusively for the experiment, MEDUSA queries were extracted from those actually submitted to the system by users.

6. Relevance information - queries derived operationally result in difficulties in providing adequate relevance judgments.

7. Size - document sets range from 500 to 50,000; request sets from 50 to 200.

8. Machine formats - differences in format inhibits the use of certain collections.

Many of these differences are related to the document collections themselves. It is for this reason that this study of the nature of document collections has been undertaken.

2.4.3

In order to combat the lack of comparability and also avoid the difficulties of setting up a new test collection, more recent experiments have made use of existing test collections, of which Cranfield 2 is by far the most popular. However, although solving some of the problems mentioned above, further complications arise from the use of these "standard" collections. Often they are used for purposes for which they were not designed, therefore calling the validity of the experimental results into question.

A further development has seen the use of more than one collection in an experiment, particularly the SMART Project. However, the SMART collections typically contain less than 500 documents. This has also led to the situation where conflicting results are obtained with different collections (Sparck Jones (1973)).

The ubiquitous Cranfield 2 collection is, as mentioned earlier, much used in retrieval experiments. However, it should be noted that this collection has an average of thirty terms per document. This is much higher than most operational collections and indeed other test collections. For this reason, some retrieval methods perform substantially better in this collection than in others - a point worth considering when evaluating results.

2.4.4 The "Ideal" Test Collection.

A solution to these problems would be of great benefit to IR research. To this end, Sparck Jones and van Rijsbergen (1975) proposed the setting up of an "ideal" test collection, a portable collection, available as a general purpose research tool. Its suggested characteristics are that it should be large and both various and homogeneous in the following:- content, source

type, origin, time and language. This is to be achieved by the ability of the main collection to be sub-divided into smaller collections. As far as size is concerned, the recommendations state that collections containing less than 500 documents (75 queries) are of no value in experiments, 1000 to 2000 documents (250 queries) are acceptable for some purposes, and over 10000 documents (1000 queries) may be necessary in some cases. Despite its obvious attractions, especially in terms of experiment comparability, the ideal test collection is not yet in existence.

2.5 TEST COLLECTION SIZE.

An important feature of the ideal test collection is its size. It was proposed that two collections be constructed, one in science, one in arts, each containing 30,000 documents, much larger than existing test collections. As stated in 2.3.2, test collections contain only a fraction of the number of documents in commercial databases, yet the underlying assumption governing their use in IR experiments is that any results will be reproduced when the system under examination is used operationally with a full database. This is known as the "same difference" principle (Robertson, 1975). Thus, in

order to obtain meaningful results, the test collection should be representative of the operational data, that is, it should exhibit similar characteristics. It can be argued that collections of 500 documents are not representative of anything, let alone a collection of 1 million documents.

2.5.1

An illustration of the problem of extrapolating results from experiments using small test collections is given by Gibbs (1977). Experiments investigating the discrimination value method (Salton, Yang and Yu, 1975) using a collection of 45,000 MEDLARS documents instead of Salton's MED450 collection (450 documents), showed that claims of improved retrieval performance were unsubstantiated, (The two collections were indexed differently - manually assigned MeSH terms in the case of the Gibbs collection, whilst MED450 was indexed automatically by SMART methods - but this was thought to have little effect).

2.6 AIMS OF THIS RESEARCH.

The aim of this thesis is to investigate the effects of size on a document collection (MEDLARS), in an attempt to determine the size of collection necessary to reflect the nature of the full collection, and yet remain small enough to handle in retrieval experiments.

In order to achieve this, subsets of varying sizes chosen merely on the basis of chronological order from the MEDLARS document collection are examined in terms of the following characteristics :- the behaviour of single term and combinations and the clustering structure.

The perhaps predictable outcome of this research is that the change is gradual as collection size increases, and that no clear cut-off point is apparent. Intuitively, however, very small collections (< 500 documents) are not representative for IR experiments.

3 EXPERIMENTS ON THE MEDLARS DATABASE.

3.1 INTRODUCTION

In this chapter, the characteristics of collections containing varying numbers of MEDLARS documents are investigated in an attempt to determine the size of collection necessary to reflect the behaviour of a full MEDLARS collection. MEDLARS has been chosen because it is indexed manually using a controlled vocabulary of index terms. The use of an automatically indexed collection may have introduced undesirable properties peculiar to the indexing method and not due to the characteristics of the collection itself. Thus, by using a manually indexed collection it is hoped that any results will hold good for other collections.

The notion of size in IR is ambiguous and requires clarification. The primary meaning here is the size of collection, that is the number of documents a collection contains. However, the number of terms in the vocabulary, the dictionary size, is also of interest.

3.2 DATABASE SIZE - THEORETICAL ASPECTS

In determining the effects of size on the function of an IR database, it is important to consider some theoretical aspects of collection size before attempting any practical experiments.

3.2.1

One of the fundamental tasks in IR is the representation of the document. This is achieved by combining various facets to produce a document description. One way of looking at this is Salton's vector representation. A document is represented by a vector of length n, each component of which corresponds to a term (n terms in the dictionary). A '1' indicates that the term is present in the description and conversely, '0' indicates that it is absent. Using this vector notation, a document can be viewed as a single point in an n-dimensional 'document space', and a collection as a set of points in the document space.

3.2.2

Given that the number of terms in the dictionary is known, a number of measures can be used to calculate the size of collection that dictionary can represent

If there are n terms in the dictionary, then the largest number of documents which can be uniquely represented, D, is given by:

1. if all n terms may be used to index a document

$$D = 2^n - 1$$

In this case, every possible point in the document space represents a document and is therefore the absolute maximum number of documents. Whilst it is theoretically possible that a document can be indexed by all the terms in the dictionary, e.g. an extensive review article, it is usually a much smaller number of terms that are assigned to any one document.

2. if the number of terms that can be applied to a document is restricted to exactly m terms

$$D = nC_m$$
$$= n!/(m!(n-m)!)$$

Again, this is unlikely because not every document will be indexed by the same number of terms.

3. if a document is indexed by at most m terms

$$D = \sum_{r=0}^{m} nC_r$$

$$= \sum_{r=0}^{m} n!/(r!(n-r)!)$$

This is the more realistic alternative. m is the maximum number of terms assigned to a document in the collection.

3.2.3

The full impact of these measures is not apparent unless some actual figures are introduced.

As an example, suppose a dictionary has 50 terms (n=50)

1. from equation 1 - all n terms may be used to index a document

$$D = 2^n - 1$$
$$= 2^{50} - 1$$
$$= 1.12 * 10^{15}$$

By comparison, MEDLARS contains $2 * 10^6$ documents from 10,000 terms and the Cranfield 1400 collection has 1,400 documents and 2,683 terms.

2. Exactly 20 terms are used to index every document (m=20) Equation 2 gives

$$D = nCm$$
$$= 50\ C\ 20$$
$$= 4.71 * 10^{13}$$

3. A maximum of 20 terms can be applied to a document - from equation 3

$$D = \sum_{r=0}^{m} nCr$$

$$= 1.14 * 10'14$$

The dictionary in these examples is deliberately constrained to an artificially small number of terms. This is to ensure that the figures remain at least within the bounds of imagination. Even so, the numbers of documents that can be represented by such a small number of terms are huge. For a dictionary containing a more realistic number of terms (say, 10,000), the figures are verging on the infinite (2'10000).

The number of documents in the MEDLARS database (2*10'6) is only a minute fraction of this theoretical maximum and yet uses the same number of terms (10,000).

The main Cranfield 2 collection has 2,683 terms and yet only 1400 documents, an even smaller proportion of the maximum possible.

An initial conclusion that may be drawn from these figures is that the dictionary sizes are too large. Indeed, working in reverse, the CA Condensates database of over two million documents could be uniquely represented by only 21 terms. The variety generation method suggested by Lynch (1977), which uses word fragments (digrams and trigrams) instead of keywords to index documents, has succeeded in reducing the total number of tokens necessary for indexing, although it remains in excess of the theoretical minimum.

Variety generation techniques, however, result in an increased number of "false drops", that is, documents are retrieved which are not relevant to the request. This is due to the fact that the use of word fragments necessarily destroys any meaning that is attached to the word from which they are derived. It would seem that a desirable feature of any indexing scheme is that it should retain the meaning of terms and therefore the number of terms must exceed the minimum required by some degree.

3.3 THE DOCUMENT COLLECTION.

3.3.1 Origin.

The experiments reported in this thesis were performed on a set of documents originating from the MEDUSA project (Barker et al (1972), Barraclough et al (1975)) which used selected documents from the National Library of Medicine (NLM) MEDLARS citation files. The sole criterion for selection was that the original article was written in English.

3.3.2 Brief description of MEDLARS.

MEDLARS documents are indexed manually at NLM using a controlled vocabulary of about 10,000 Medical Subject Headings (MeSH terms). These terms may be made more specific by the addition of certain qualifiers of which there are about 70 available, and associated with each term is a list of valid qualifiers. Within the dictionary there are 38 general purpose terms, known as "check tags" (Appendix 1). At least one of these terms must be assigned to each document and this serves to give a broad

indication of the content of the document. On average, a document is indexed by about 12 terms, qualified and unqualified, of which 3 are "print terms" under which the article appears in "Index Medicus".

MEDLARS query formulation is by means of a Boolean combination of MeSH terms using the operators AND, OR and NOT. The terms may be qualified by using a LINK to a valid qualifier.

3.3.3 Modifications and simplifications.

The MEDUSA system utilised the capabilities of the Newcastle File Handling System (NFHS) (Cox and Dews (1967)), developed specifically for the processing of bibliographic data. Whilst this proved very effective for the provision of an information service via the MEDUSA system, the experiments reported here were concerned with investigating the document representation, and as a result many fields of the citation records were superfluous. Indeed, in the majority of cases the only requirement was easy access to the index terms of the document - such fields as author, title, journal, volume and date of publication were not required. The use of the citation files in their original NFHS format would have incurred

substantial overheads in the form of increased computation time and larger storage requirements. For this reason, the index terms were extracted from the citation files and the use of NFHS routines and records was discontinued. A further simplification involved ignoring qualifiers, thus multiple occurrences of the same term with different qualifiers in the same document were treated as a single occurrence of that term. Also, no distinction was made between print terms and ordinary terms, although it could be said that print terms were more important in comparison to others. Early experiments indicated that because of their enforced use in indexing, check tags occurred with unusually high frequencies and therefore were removed from both the documents and the dictionary. This resulted in the dictionary being reduced to 10,137 terms.

The final form of the documents was similar to the 'ab' format suggested by Sparck Jones & Bates (1977) and indeed differed only in that the termination symbol "/" was omitted and the number of terms in the document was the first number in the document description to compensate for this.

a n b1 b2 b3 ... bi ... bn

a = document number: simplified to between 1 and 61036

n = number of terms in the document

bi= term number: simplified to between 1 and 10137 for ease of frequency calculation. The original terms were in the range 24 to 124908 with obvious gaps and a conversion list was available should the identity of the term be required.

3.3.4 The full and sub-collections.

The full collection was composed of a four month section of citation files, each month containing some 15,000 documents giving a total of 61,036 documents.

Sub-collections were formed containing 500; 5,000; 10,000; 15,000; 20,000 and 25,000 documents. (In addition, sub-collections of 250 and 1,000 documents were used in the clustering experiments of Chapter 4.) The 500 sub-collection comprised the first 500 documents of the first month, the 5000 sub-collection the first 5,000 documents and so on. In this way each sub-collection was a subset of the next largest sub-collection. This enabled the effects of changing document collection size to be investigated.

3.3.5 Randomness of data.

In order that the sub-collections could be realistically compared both with each other and the full collection, it was desirable that the documents should have a degree of randomness.

As mentioned earlier, only English language documents are contained in the collection. This obviates the problem of the formation of groups caused by the indexing of non-English language documents in batches.

The average number of terms per document remains reasonably constant throughout the collections (Table 3.1). This indicates that any differences between collections are not due to variations in the level of indexing, i.e. the use of more or less terms to index documents.

It is possible that a collection may contain too many documents in the same subject area and therefore randomness may be lost due to subject grouping. However, this is not the case in MEDLARS. As the indexing is a continuous process, any specialist journal, which may contain between 10 and 20 articles, will be absorbed into

Sub-collection statistics

No of Docs	No of term occurences	Av no of trms/doc	No of diff terms	Real % of terms used
500	4254	8.51	2128	20.99
5000	46308	9.26	6521	64.33
10000	94998	9.45	7596	74.93
15000	145798	9.71	8129	80.19
20000	191804	9.59	8422	83.08
25000	241914	9.68	8639	85.22
61036	572707	9.38	9189	90.65

Table 3.1.

the collection immediately and not held over until there are more in the same subject, which may result in a block of 200 or so articles in the same subject area.

Experiments on 2-term combinations, which are reported later in this chapter, have specifically included a test for randomness.

3.4 EXPERIMENTS ON SINGLE TERMS.

This section describes experiments performed on the collections investigating the behaviour of single terms, that is, terms considered independently of others contained in the document descriptions.

Single terms are examined with respect to term frequency and the total number of different terms used.

If the sub-collection is a true subset and therefore representative, there should be no variation in relative frequencies between sub-collection and full collection and all the terms in the full collection should be present in the sub-collection.

3.4.1 Term frequencies.

A useful indication of the behaviour of a document collection is the frequency with which index terms occur. For this reason, the frequencies of occurrence of each term in all the sub-collections and the full collection were obtained. Initial analysis revealed that certain terms had a surprisingly high frequency compared to others throughout all the collections. Upon examination, it was found that the terms were in fact, "check tags" (see 3.3.2), and it was assumed that no matter what size of collection was analysed these terms would occur with exaggerated high frequency and were therefore ignored.

As it was intended to draw graphs to compare the frequencies of term occurrences in each sub-collection both with each other and the full collection, it was necessary to ensure that the graphs were indeed comparable. In order to achieve this, an ordering of

terms was obtained by ranking the terms in decreasing frequency of occurence in the full collection. Terms of equal frequency were ranked in numerical order. The terms were then assigned to 100 divisions, each division containing 100 terms, so that division 1 contained the 100 most frequent terms, division 2 the next 100 frequent terms and so on.

For each division a frequency representative was calculated by summing the frequencies of the individual terms. This was then normalised by dividing by the total number of term occurences, the final figure showing the proportion of terms in that particular division. Similarly, graphs were drawn for the sub-collections, but the divisions obtained for the full collection were retained in order to avoid the problem of terms "migrating" from one division to the next making comparison impossible.

The normalisation of frequency representatives and the retention of the same divisions ensured that the graphs were directly comparable.

Figures 3.1 to 3.7 show the graphs for the collections over the complete range of divisions (1 - 100). Figures 3.8 to 3.14 show only divisions 10 - 90 for

reasons of clarity. Divisions 1 - 9 exhibited no fundamental differences between collections.

Figure3.1

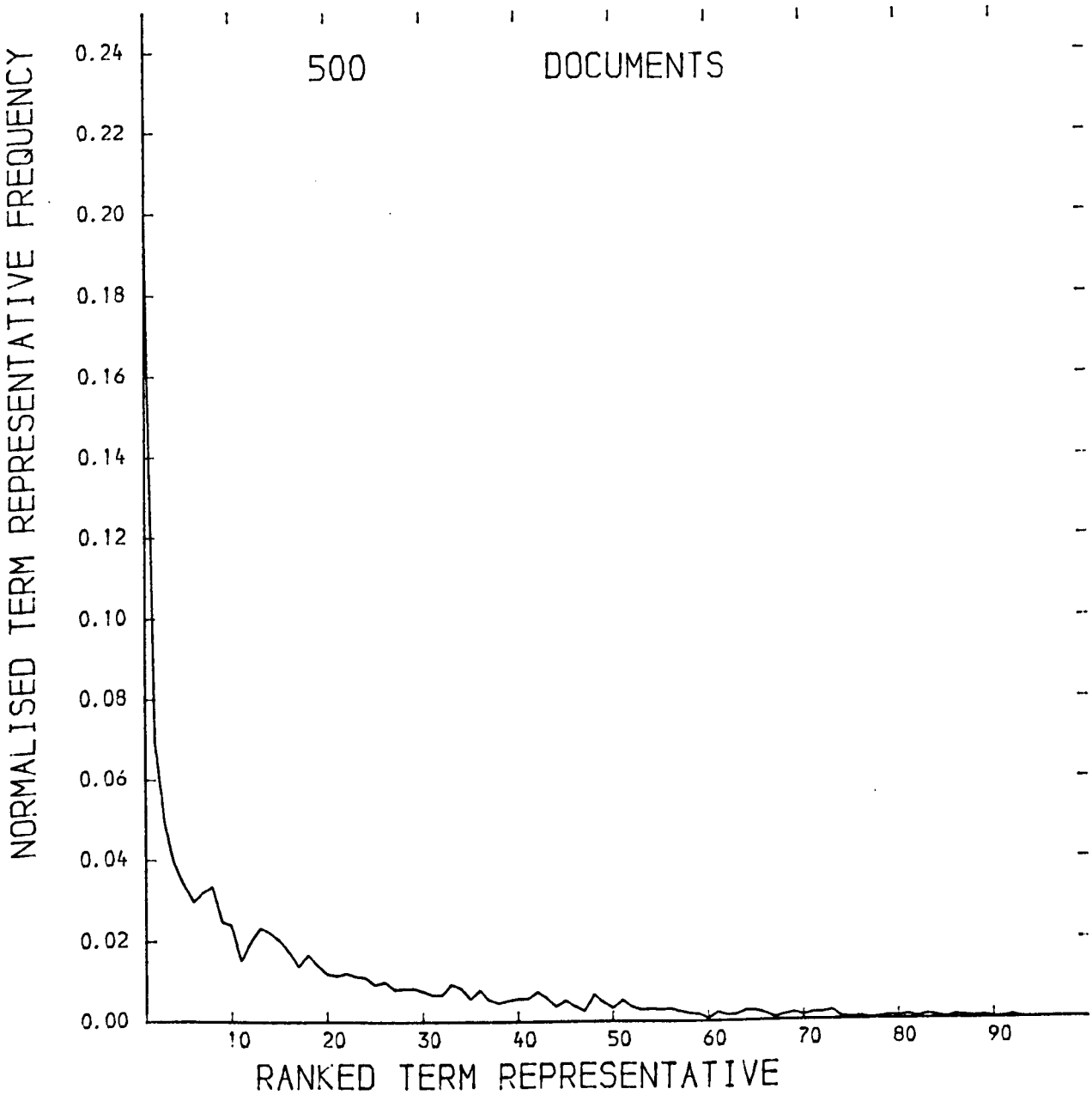

500 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10
20
30
40
50
60
70
80
90
RANKED TERM REPRESENTATIVE

Figure 3.2

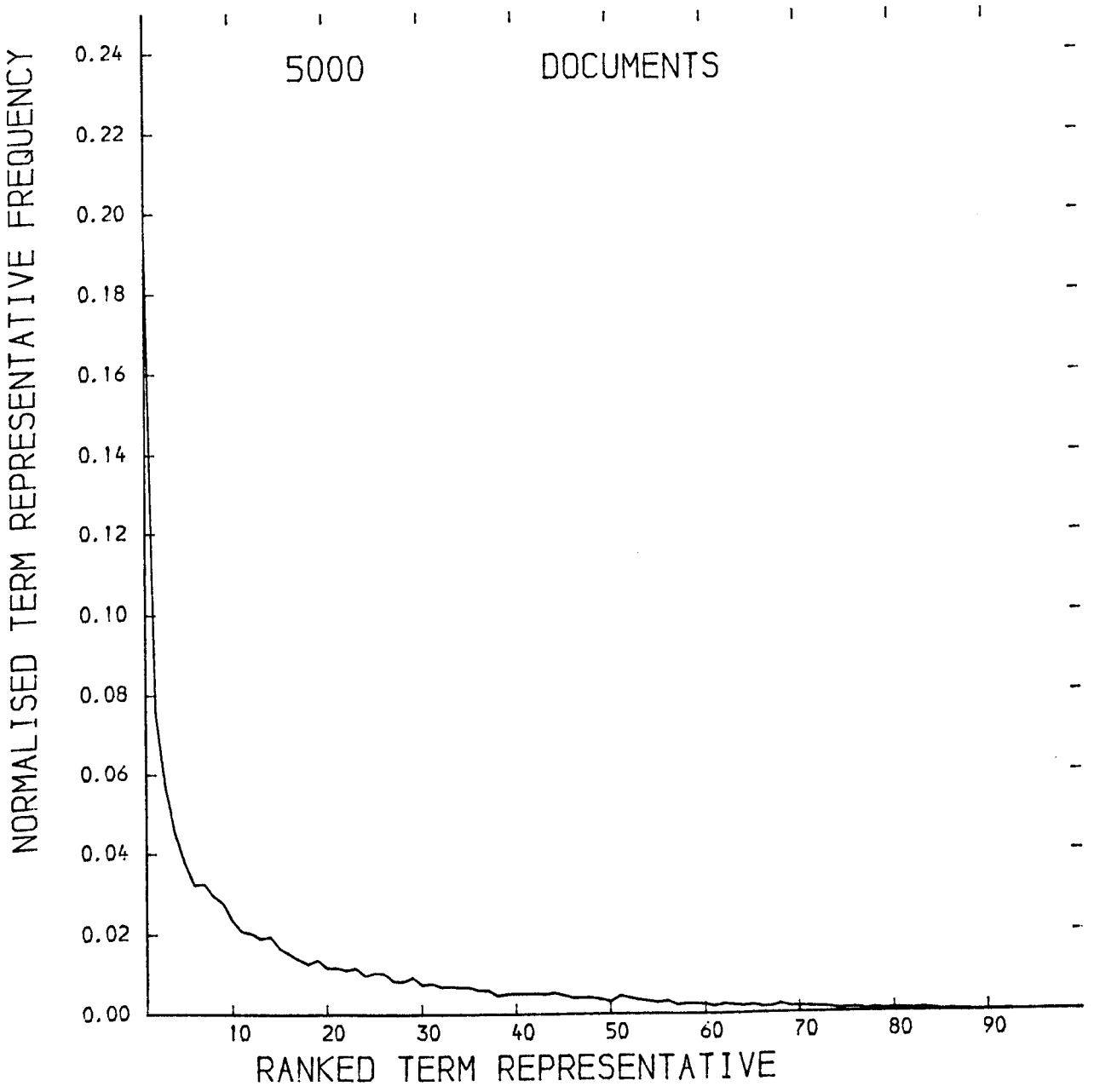

5000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10
20
30
40
50
60
70
80
90
RANKED TERM REPRESENTATIVE

Figure 3.3

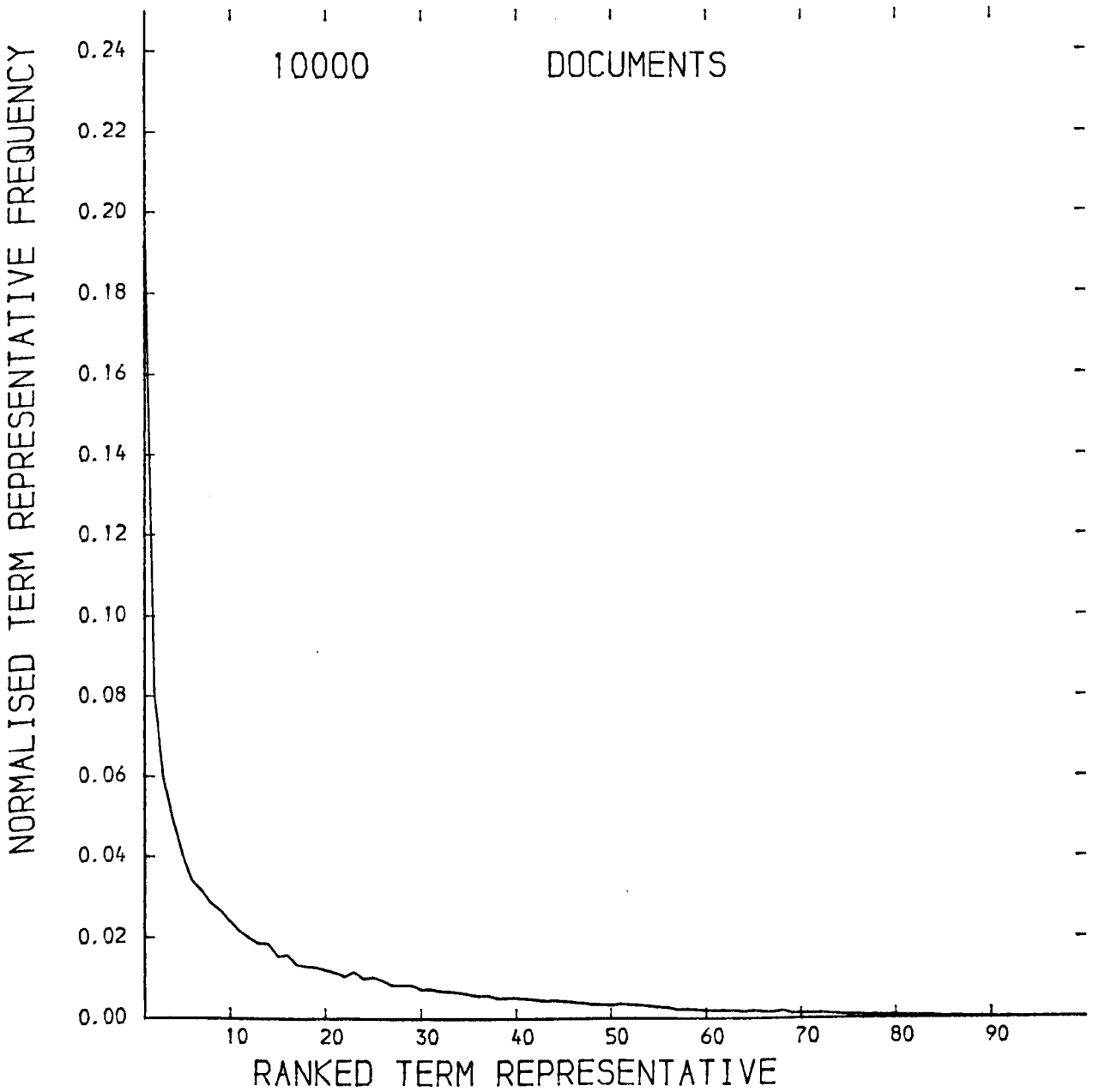

10000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10
20
30
40
50
60
70
80
90
RANKED TERM REPRESENTATIVE

Figure 3.4

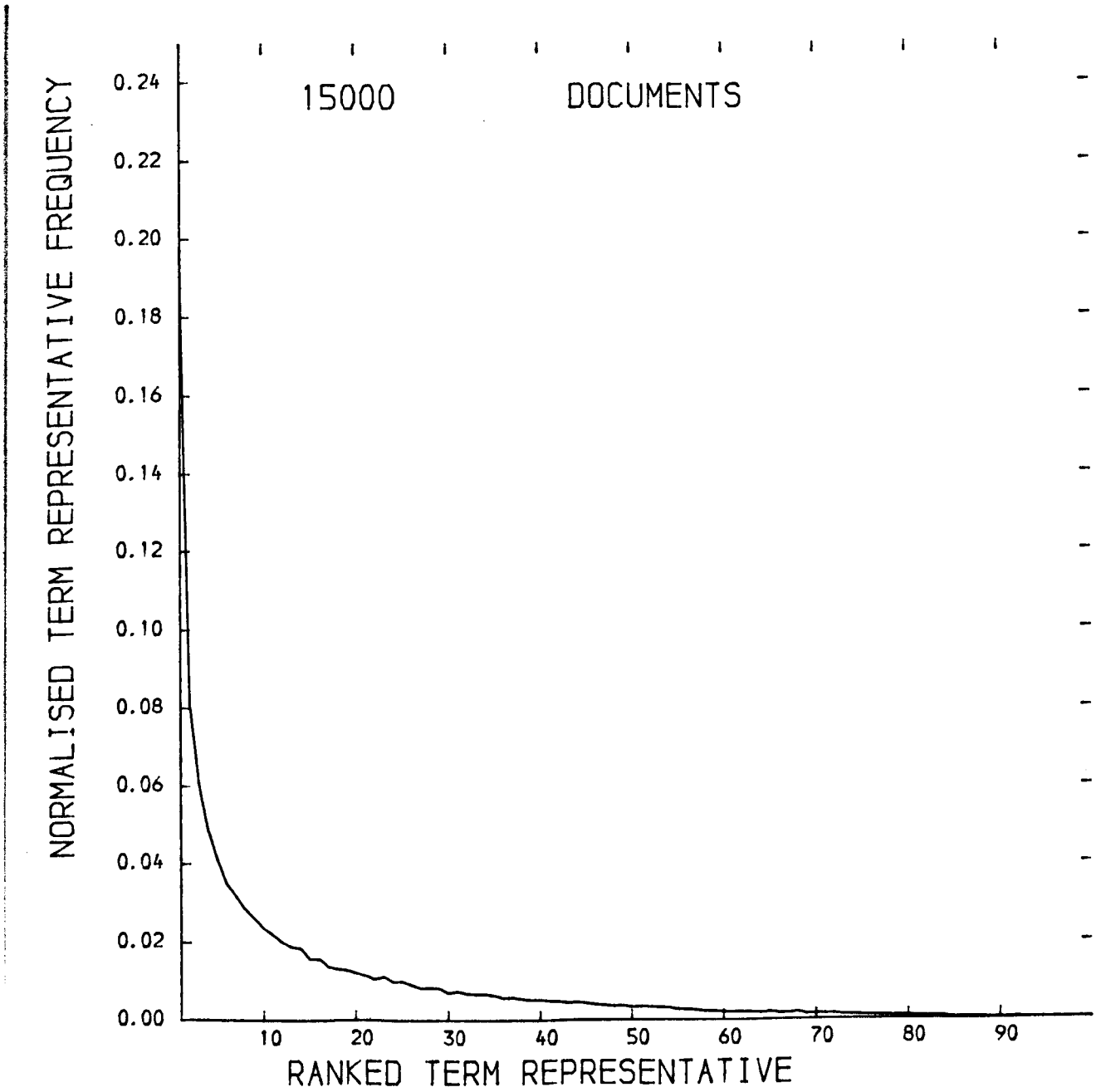

15000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10
20
30
40
50
60
70
80
90
RANKED TERM REPRESENTATIVE

Figure3.5

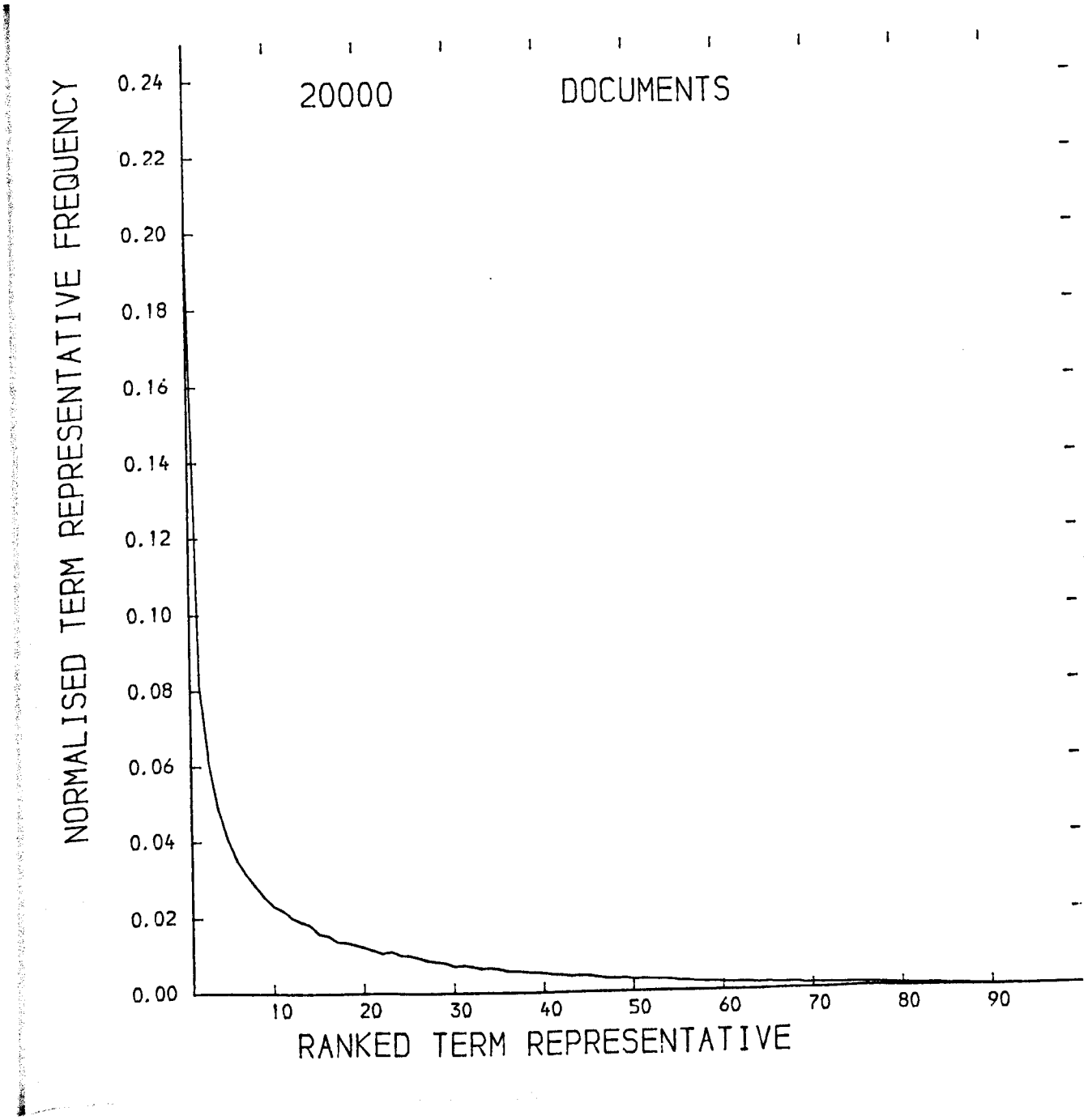

20000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10
20
30
40
50
60
70
80
90
RANKED TERM REPRESENTATIVE

Figure 3.6

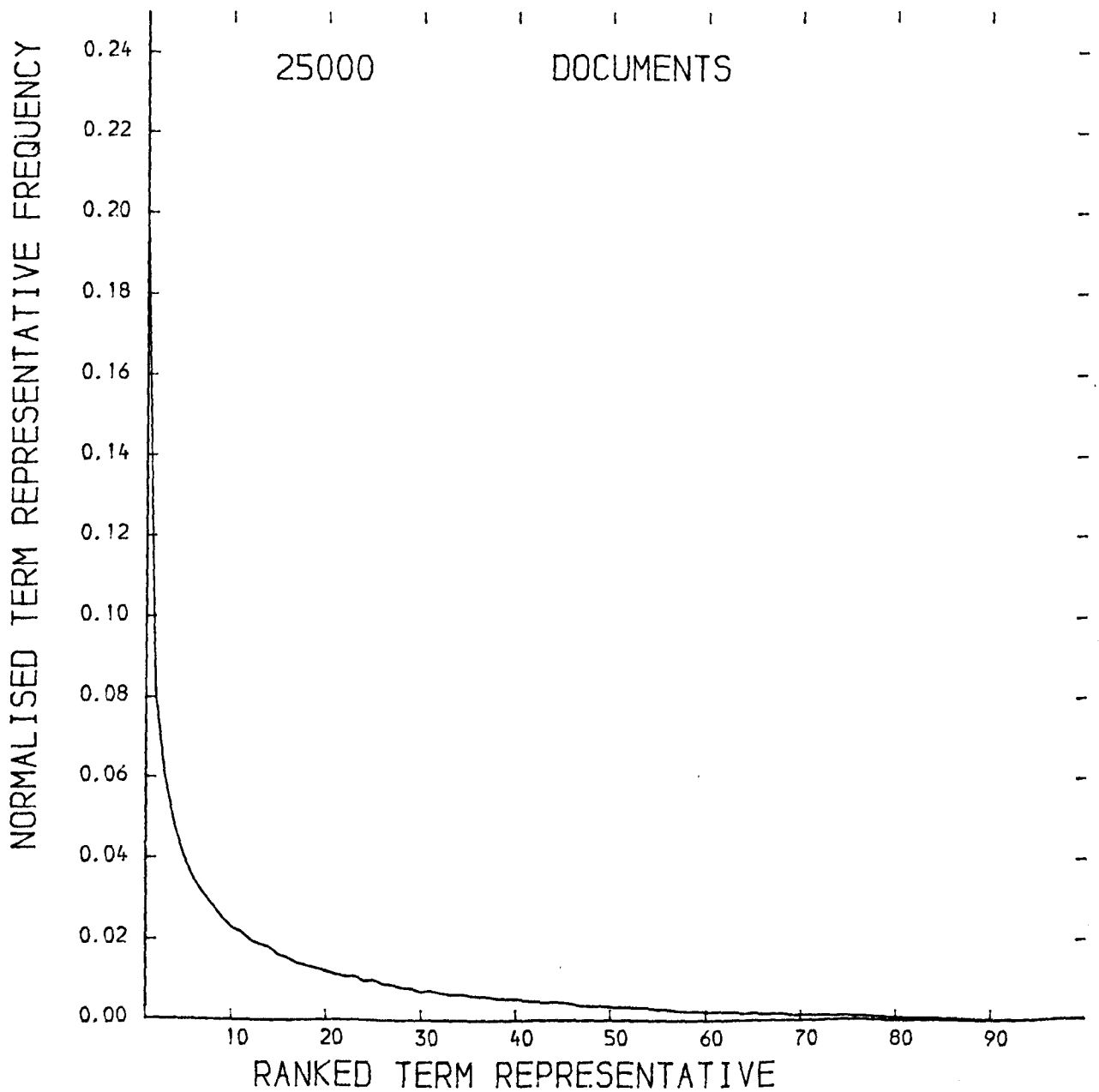

25000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10
20
30
40
50
60
70
80
90
RANKED TERM REPRESENTATIVE

Figure 3.7

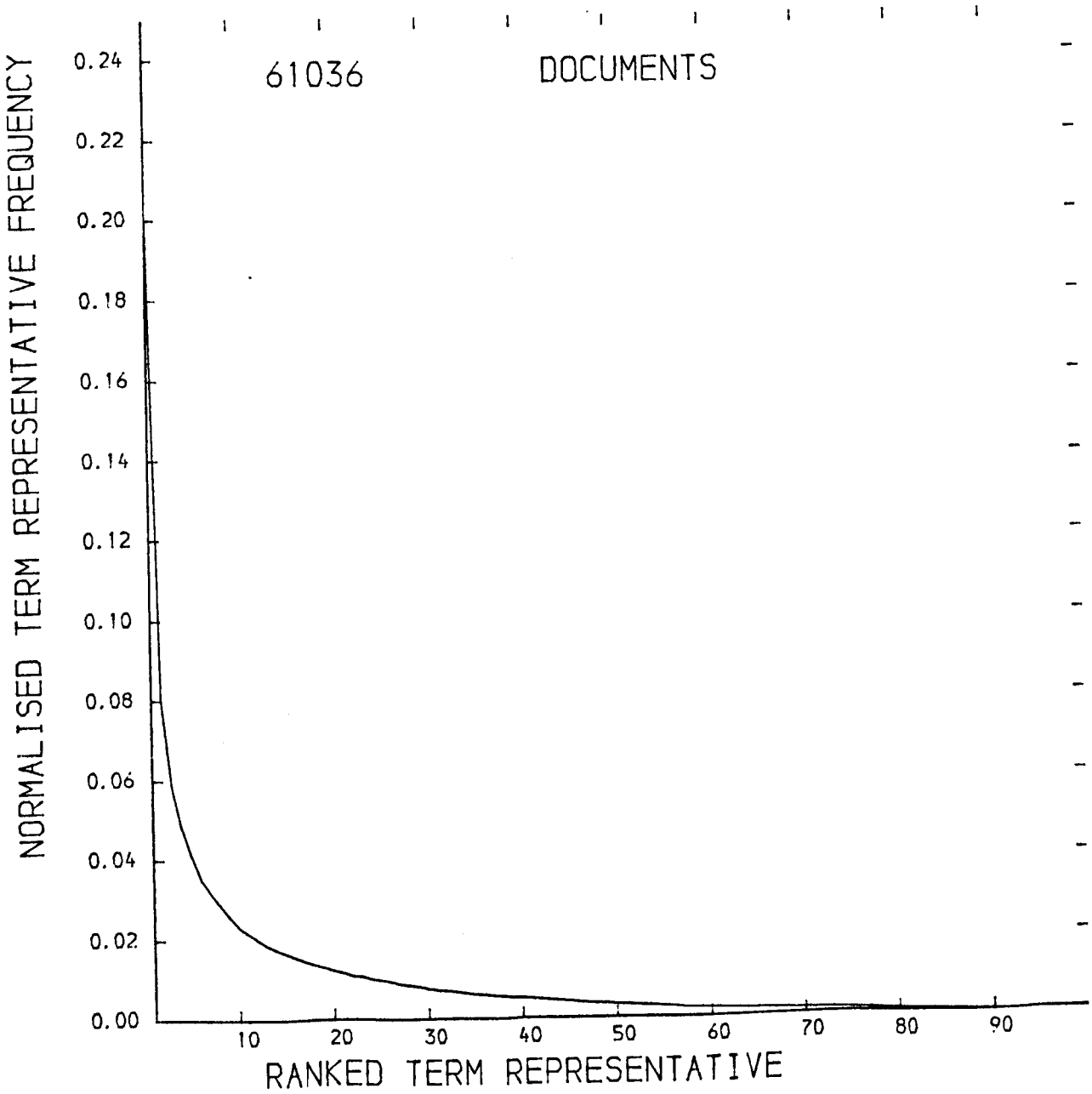

61036 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10
20
30
40
50
60
70
80
90
RANKED TERM REPRESENTATIVE

Figure 3.8

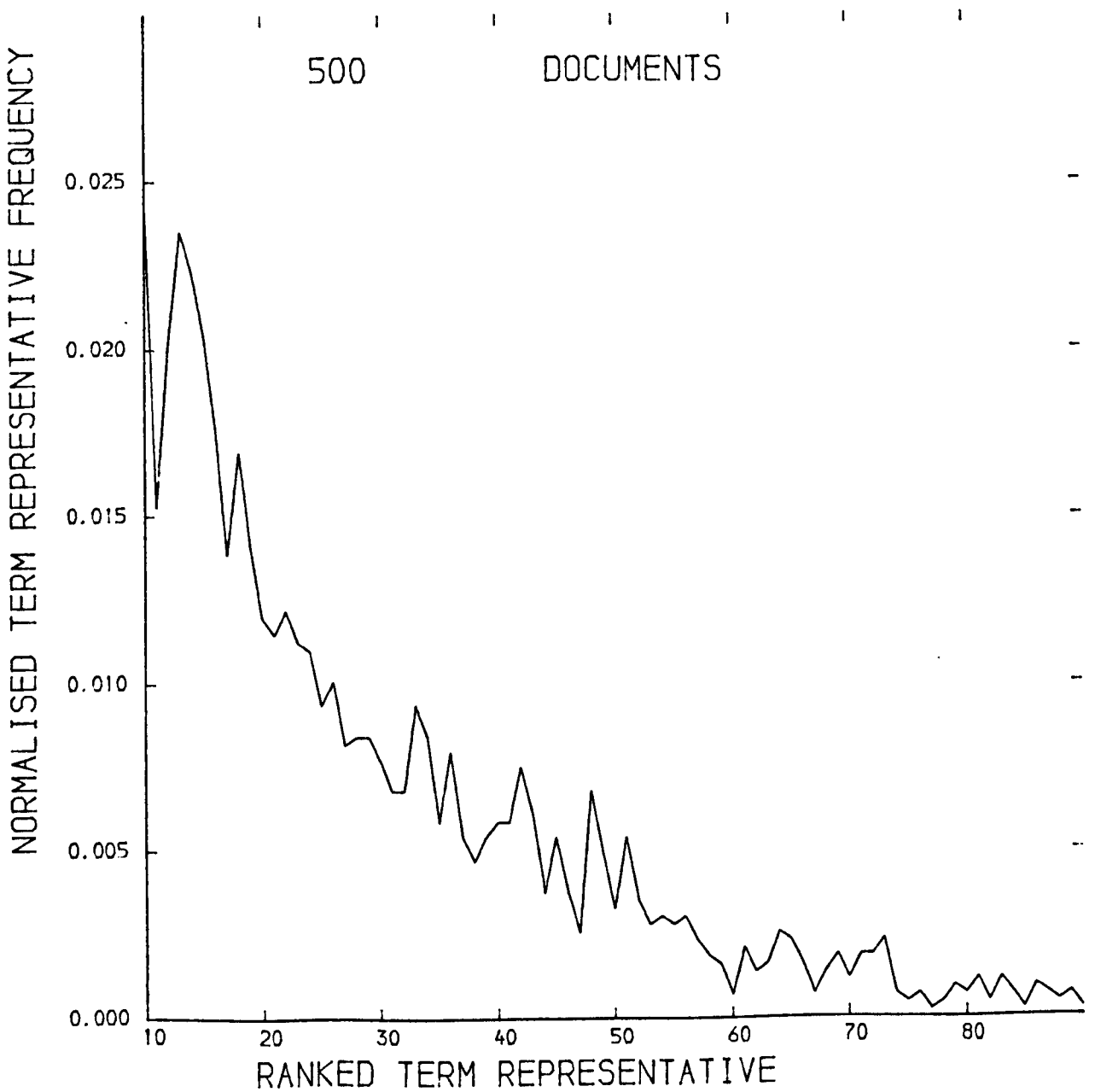

500 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
80
RANKED TERM REPRESENTATIVE

Figure 3.9

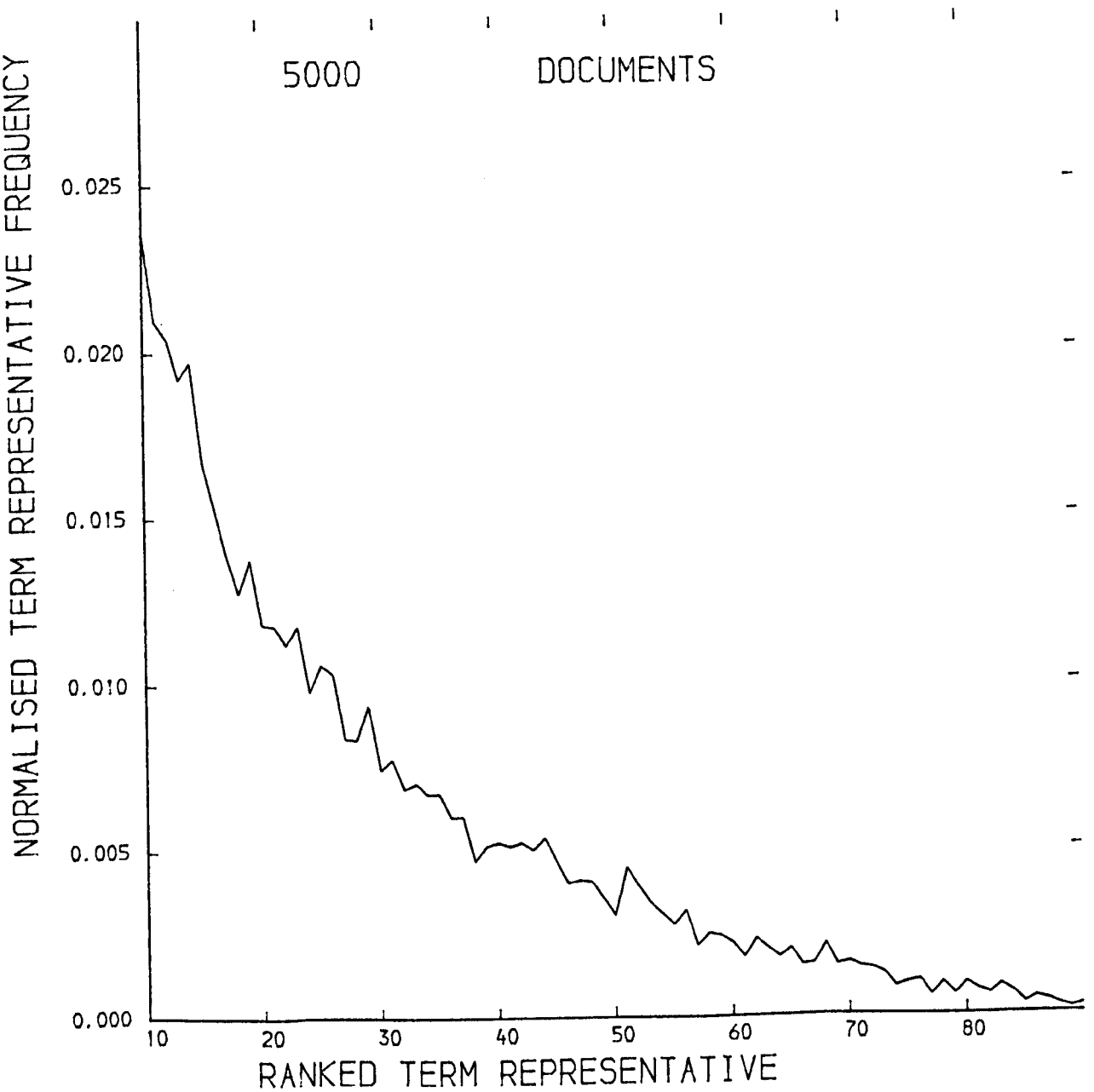

5000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
80
RANKED TERM REPRESENTATIVE

Figure 3.10

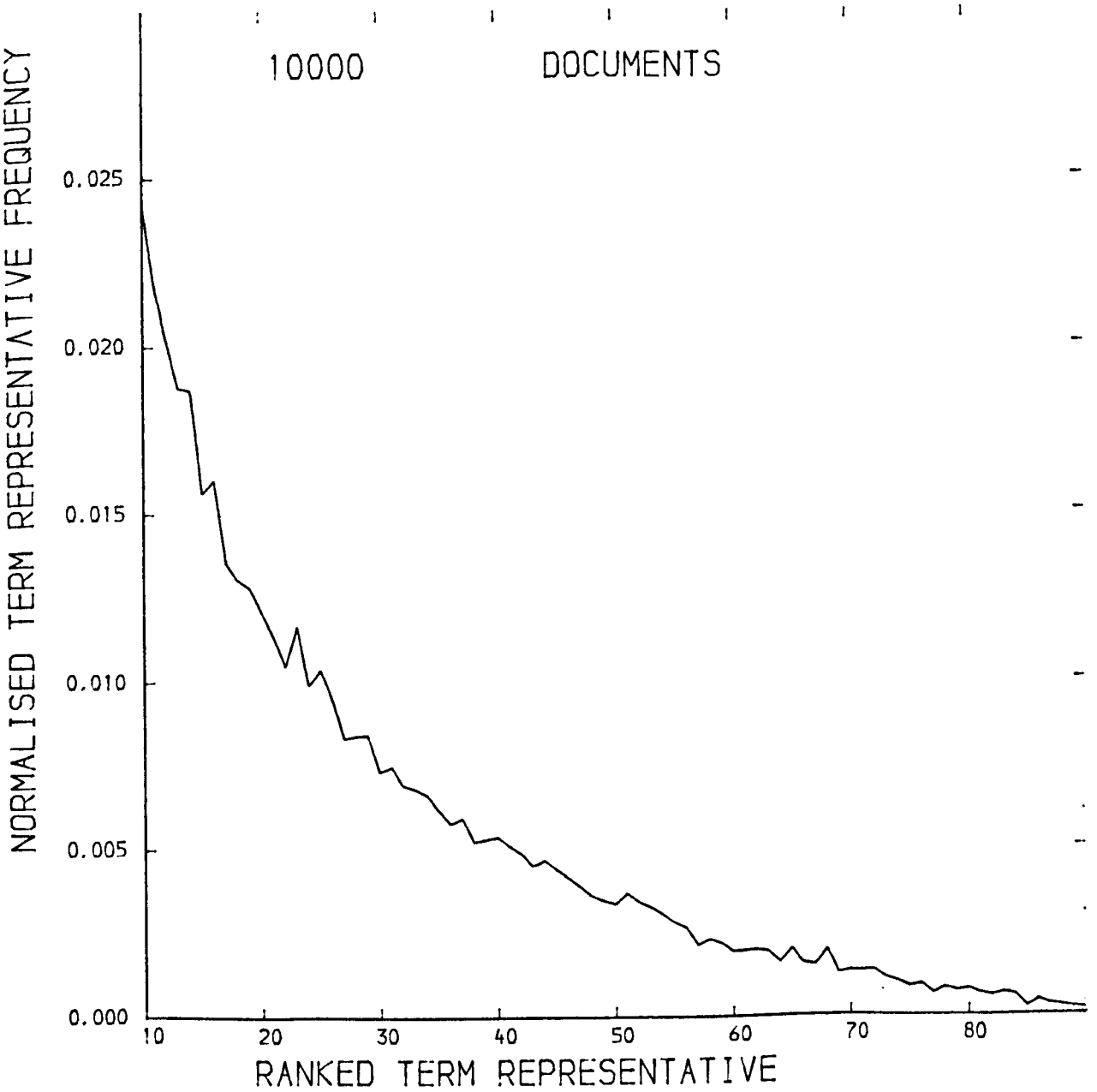

10000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
80
RANKED TERM REPRESENTATIVE

Figure 3.11

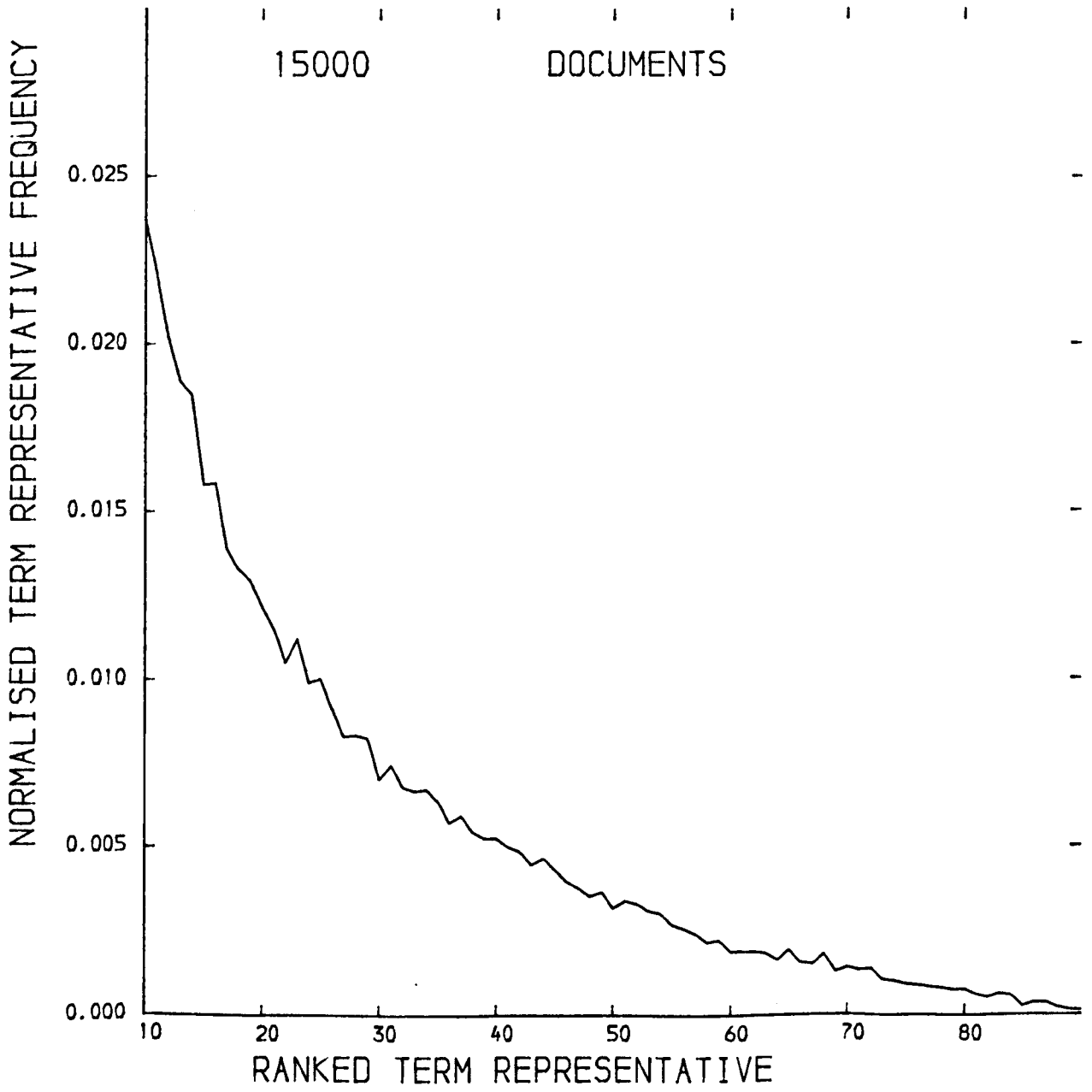

15000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
80
RANKED TERM REPRESENTATIVE

Figure 3.12

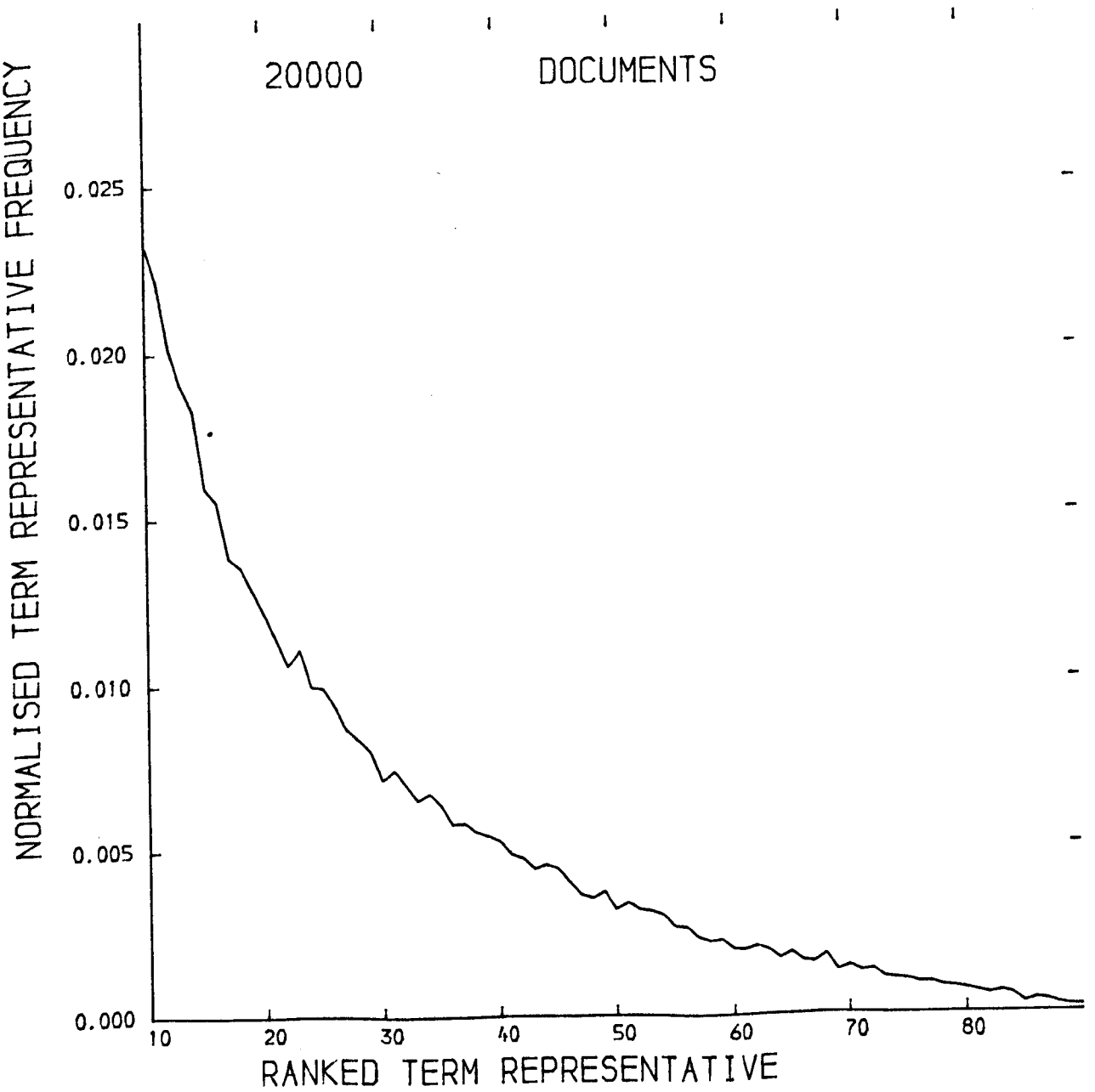

20000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
80
RANKED TERM REPRESENTATIVE

Figure 3.13

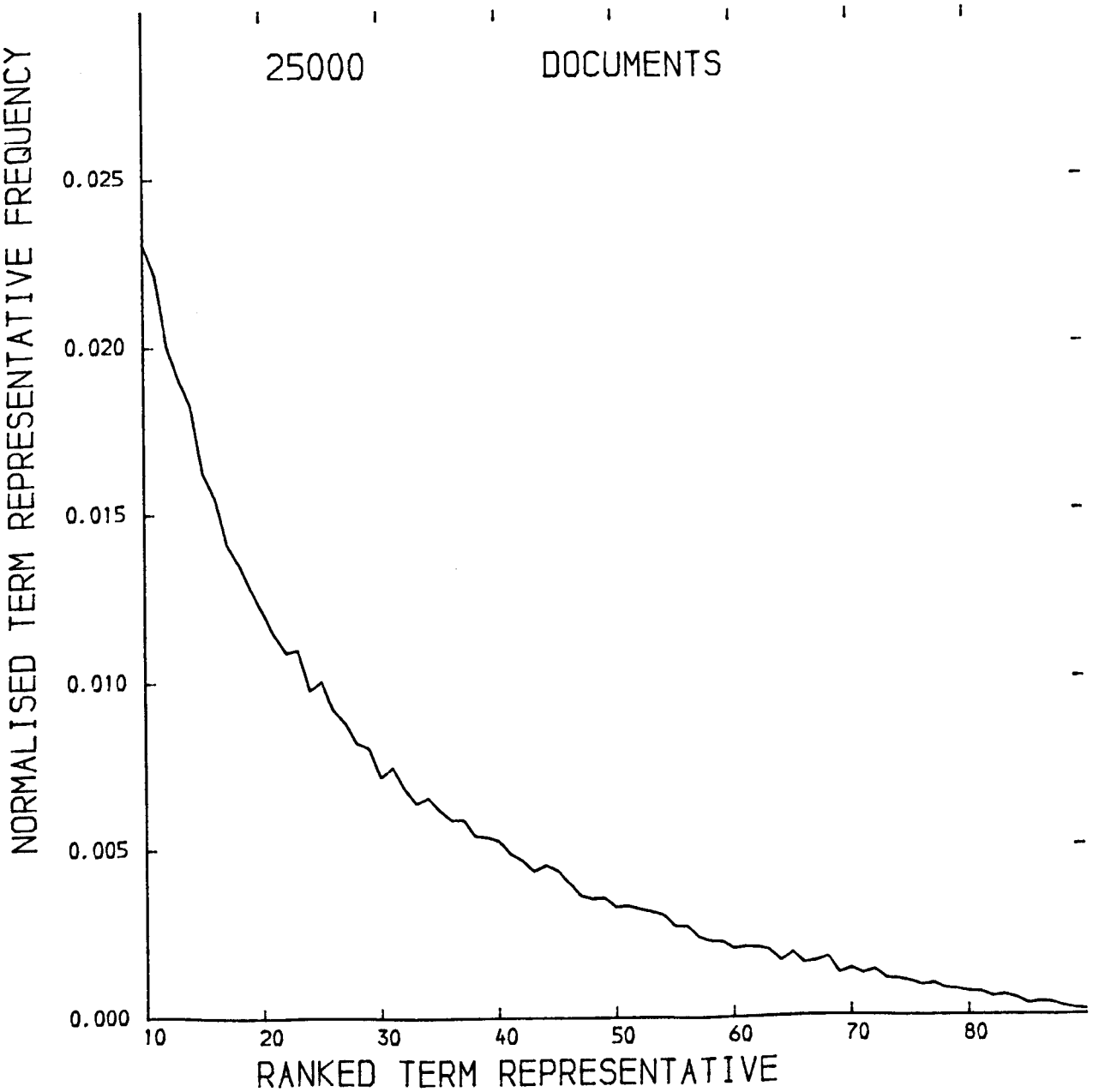
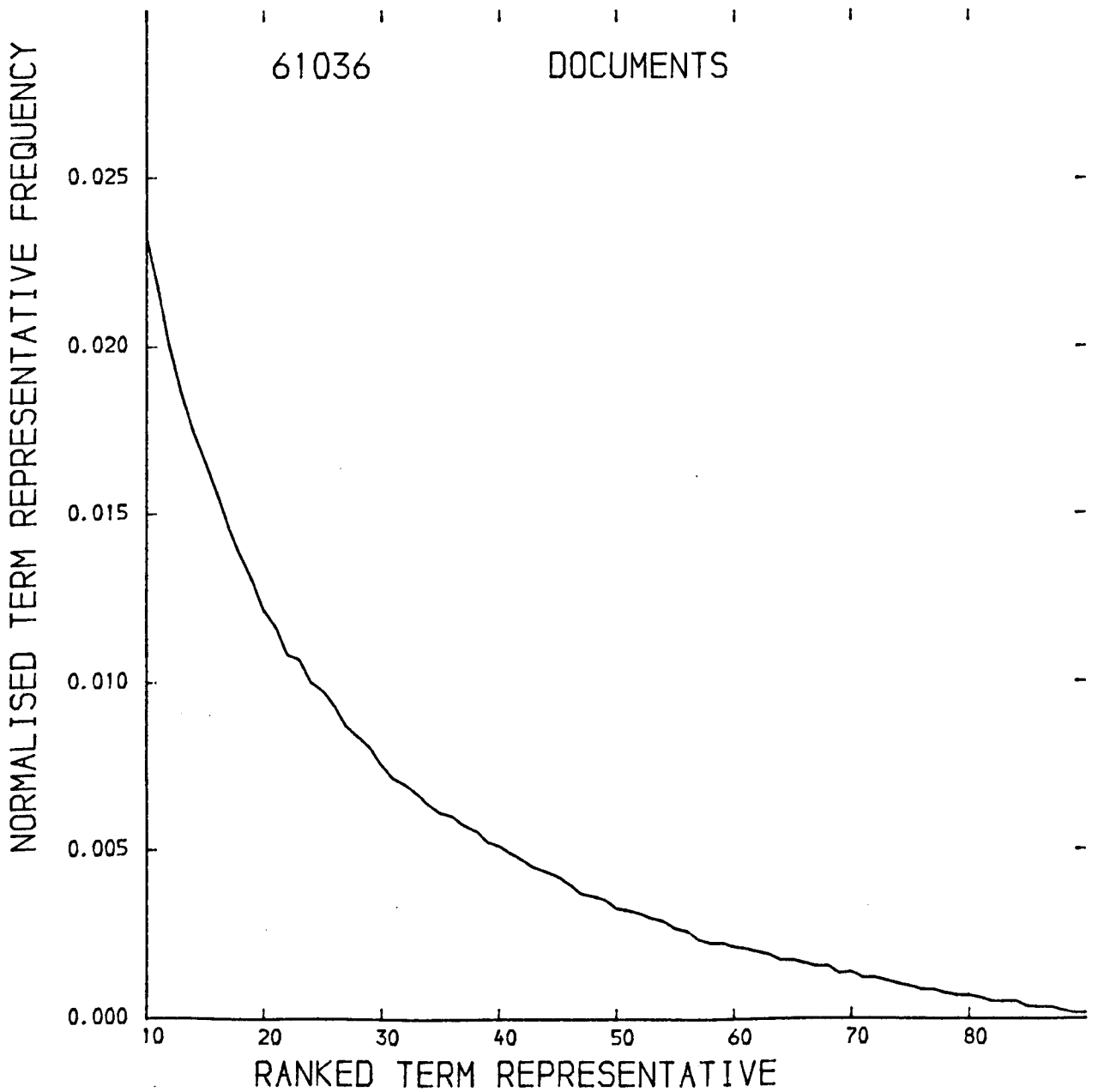

25000 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
80
RANKED TERM REPRESENTATIVE

Figure3.14


61036 DOCUMENTS
NORMALISED TERM REPRESENTATIVE FREQUENCY
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
80
RANKED TERM REPRESENTATIVE

As a result of the ordering of terms according to their frequency, the graph for the 61000 collection is by definition smooth. For the 500 collection the graph is anything but smooth with large deviations from the 61000 curve. As the collection sizes increase, the deviations become less pronounced, yet it is only in the 15000 and 20000 graphs that any sign of smoothing out can be seen. As a general trend, the peaks and troughs of the graphs do diminish with increase in size. These deviations are all the more remarkable considering the fact that the points depict only the representatives of the 100 divisions which were obtained by averaging the actual frequencis, in effect smoothing the variations within the division. Thus, graphs showing each term individually (Figure 3.15 to 3.18), reveal even greater differences.

Figure 3.15

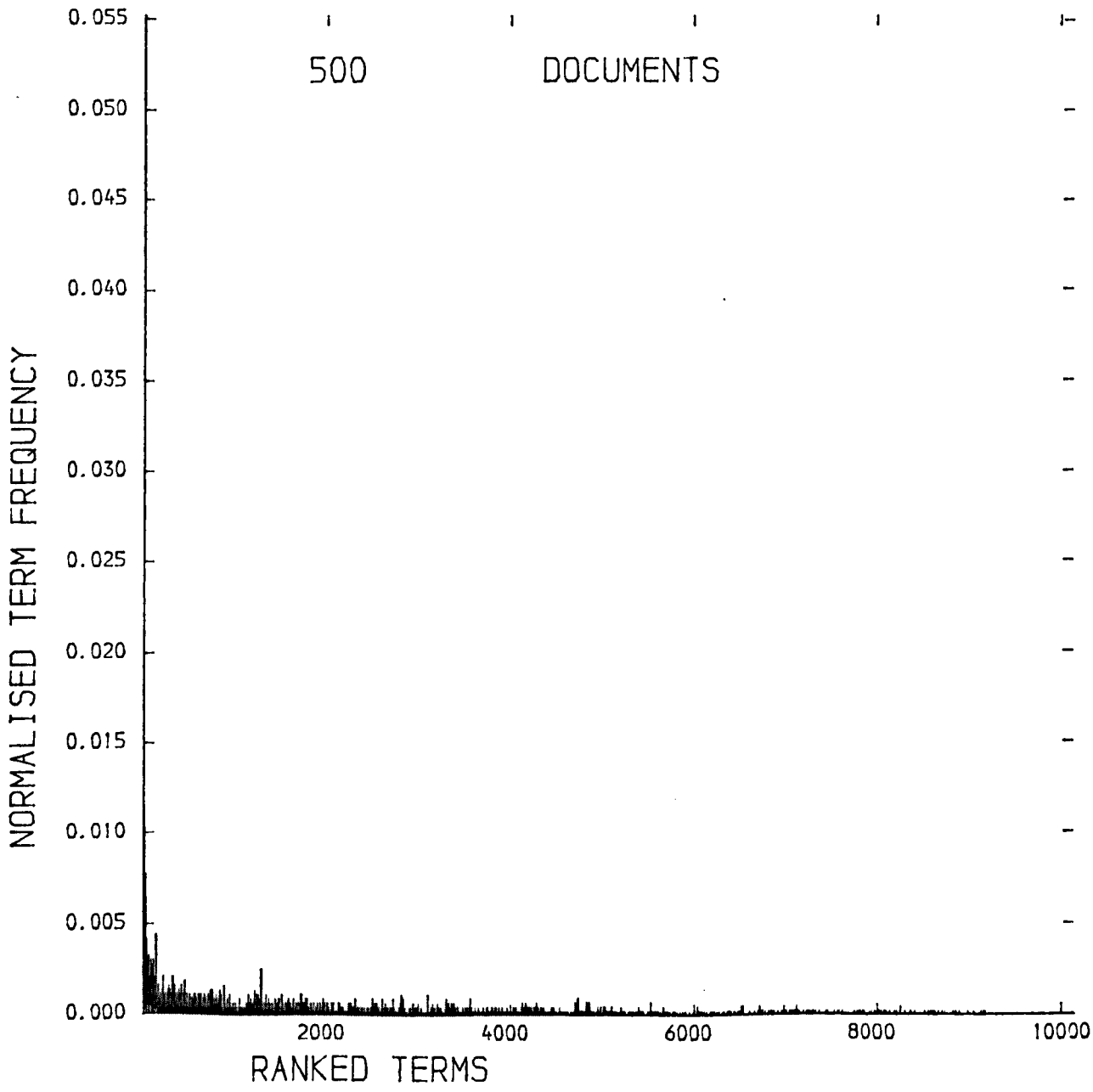

500 DOCUMENTS
NORMALISED TERM FREQUENCY
0.055
0.050
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
2000
4000
6000
8000
10000
RANKED TERMS

Figure 3.16

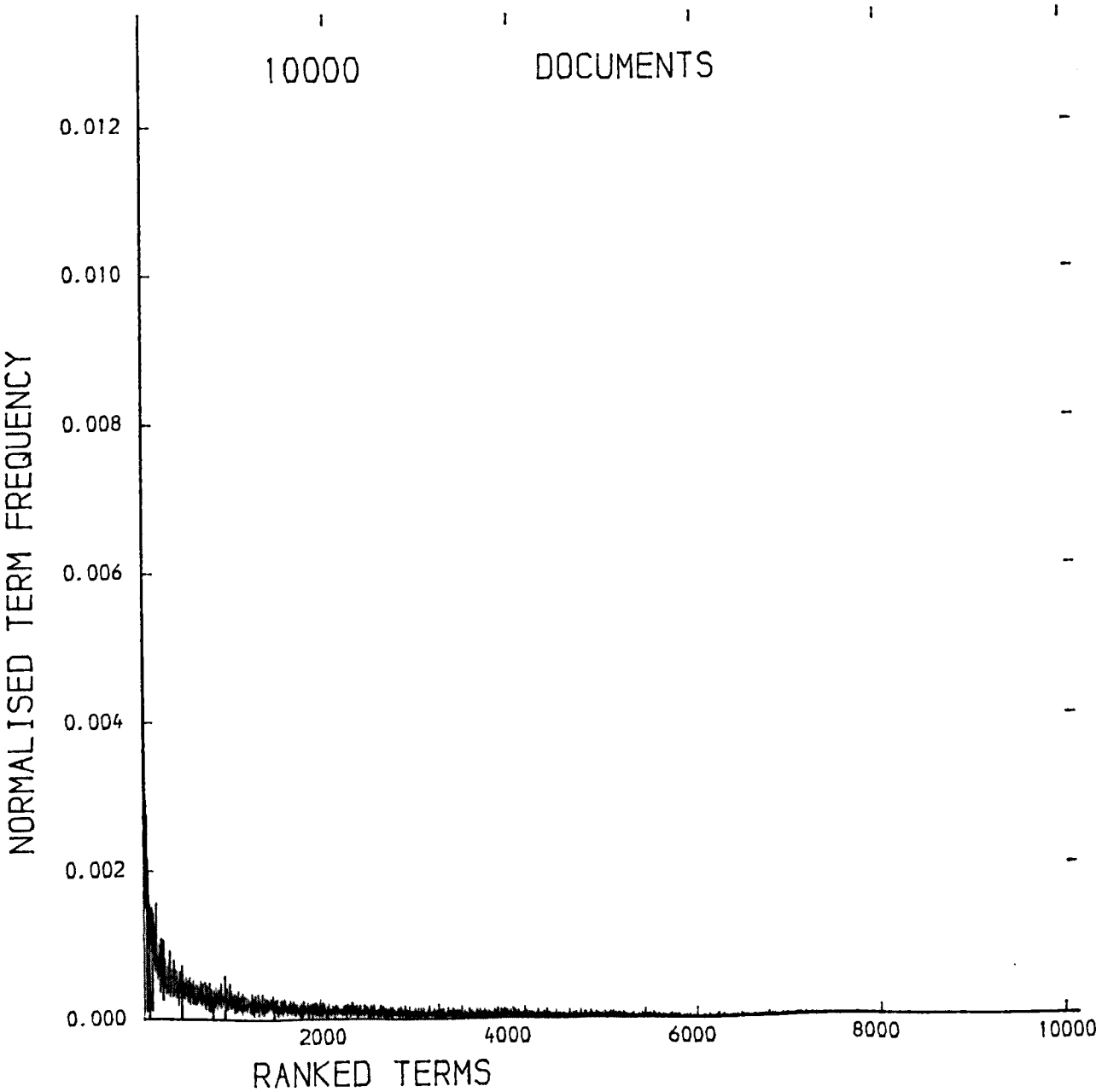

10000 DOCUMENTS
NORMALISED TERM FREQUENCY
0.012
0.010
0.008
0.006
0.004
0.002
0.000
2000
4000
6000
8000
10000
RANKED TERMS

Figure 3.17

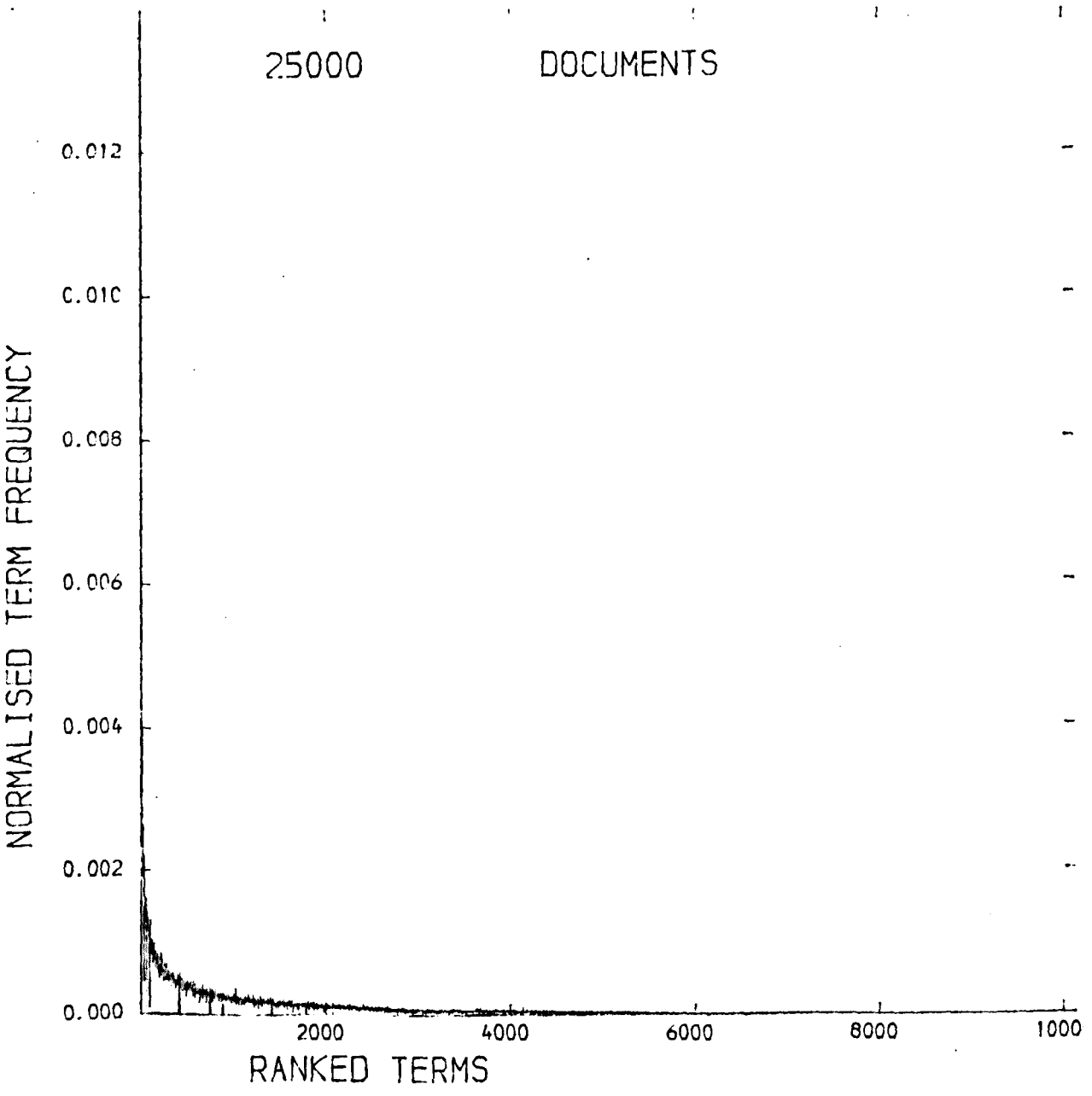
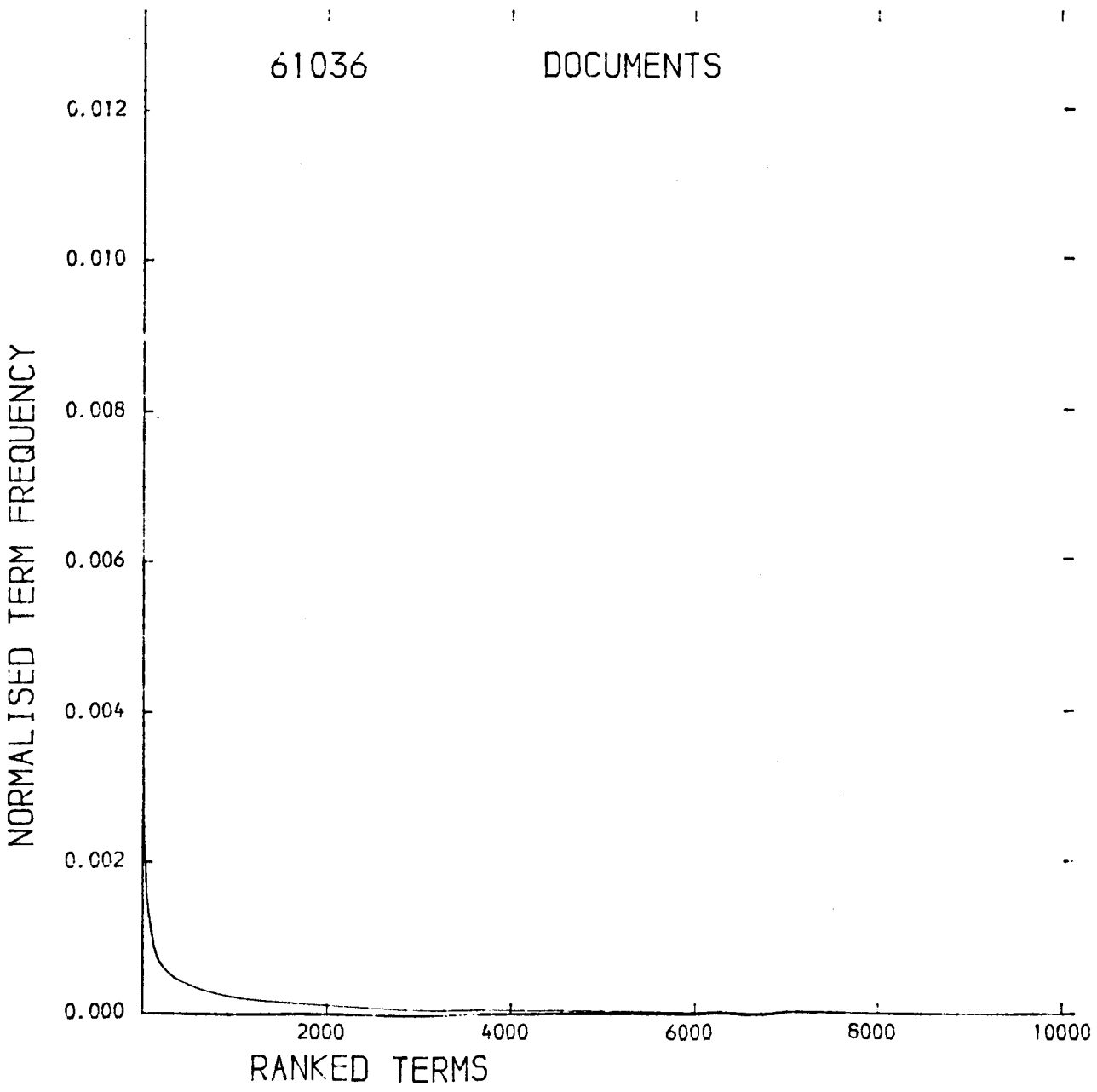

25000 DOCUMENTS
NORMALISED TERM FREQUENCY
0.012
0.010
0.008
0.006
0.004
0.002
0.000
2000
4000
6000
8000
1000
RANKED TERMS

Figure 3.18


61036 DOCUMENTS
0.012
0.010
0.008
0.006
0.004
0.002
0.000
NORMALISED TERM FREQUENCY
2000
4000
6000
8000
10000
RANKED TERMS

3.4.2 The introduction of new terms.

If a dictionary contains a finite number of index terms, then the number of different terms used in a document collection can give an indication of the coverage of the subject area. Further, a collection which, upon the addition of further documents, shows little or no increase in the number of different terms used may be deemed representative of the full collection.

Figure3.19

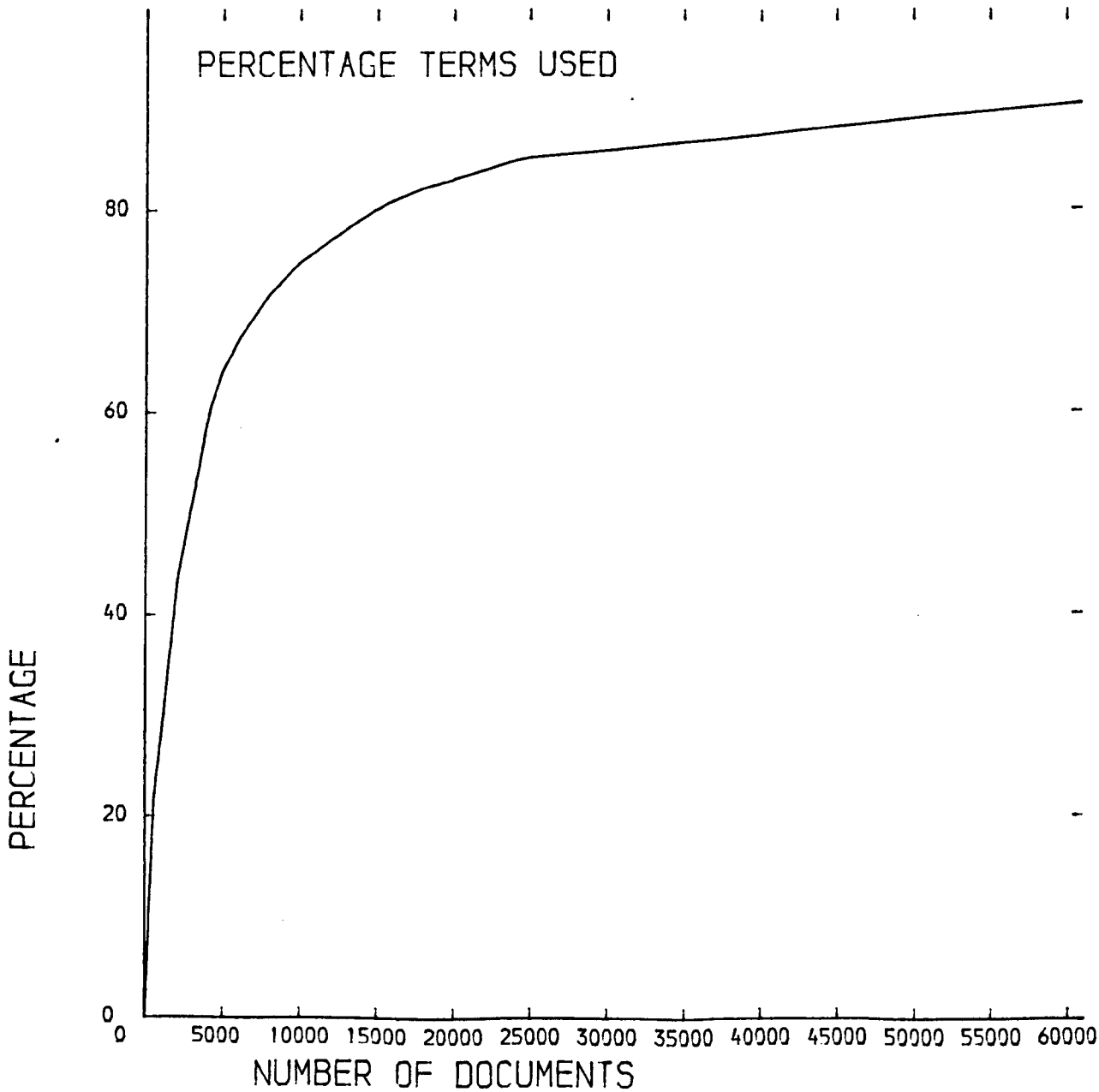

PERCENTAGE TERMS USED
PERCENTAGE
0
20
40
60
80
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
50000
55000
60000
NUMBER OF DOCUMENTS

Figure 3.19 and Table 3.1 give details of the increases in the percentage of new terms introduced in proportion to the increased size of document collection. In addition to the sub-collections, the file boundary sizes were also obtained and plotted. In a study of dictionary size, Houston and Wall (1964) maintain that if less than half the terms in a dictionary occur in the collection it is being used to index, ther the index is of little use. Turning this around, a collection which uses less than half of the terms available in the dictionary may be of limited use in retrieval experiments. In this case, the 500 sub-collection suffers from this deficiency. Indeed, it is only when around 3000 documents are used that the 50% mark is passed.

By 10000 documents almost 75% of the terms have been used and after this point there is a levelling off in the rate of increase. Indeed, it should be noted that the full collection only uses 90% of the terms available. This is a large enough percentage to deem 61,000 documents acceptable as a full collection.

The conclusions to be drawn from these initial experiments are:-

1. Collections of less than 3000 are unlikely to be

useful for experimental purposes.

2. 15,000 documents may be sufficient to be able to obtain meaningful results.

3. It is more likely that larger collections are necessary to ensure that they are representative of the full collection.

3.5 TERM CO-OCCURRENCES.

In IR, terms rarely occur in isolation. Indeed, it is the very co-occurrence of terms in documents and queries that attempts to express the content of the article through indexing and the information need of the user by the combination of terms in the search formulation.

This section considers the properties of the co-occurrence of two terms in the same document, hereafter referred to as "combinations". Of course, it is possible to consider co-occurrences of three or more terms, but it is usually the case that a query can be reduced to a combination of two basic components, linked together, for example, by AND in a Boolean query formulation scheme. Any more than two terms so linked reduces considerably the

number of documents that may be selected in response to the query. An exception to this is when one of the terms is a check tag. This provides a general mechanism for excluding documents that lie outside the particular field of interest.

3.5.1 A Theoretical View of Combinations.

In the same way that a given size of dictionary can give rise to a theoretical maximum number of documents (3.2.2), a finite number of two term combinations can be generated.

For a dictionary of 10000 terms (n=10000), the maximum number of different combinations is given by

$$C = nC2 = n!/(2!(n-2)!)$$
$$= (n(n-1))/2$$
$$= 5.0 * 10^7 \quad (50 \text{ million})$$

The actual number of combinations can be expected to be much less than this because some terms may not occur together. This is due to the basic MEDLARS indexing rule of using the most specific term available and the

hierarchy of terms within the thesaurus. For example, the following hierachy of terms is defined by MEDIARS (from "The Principles of MEDLARS", NIM):-

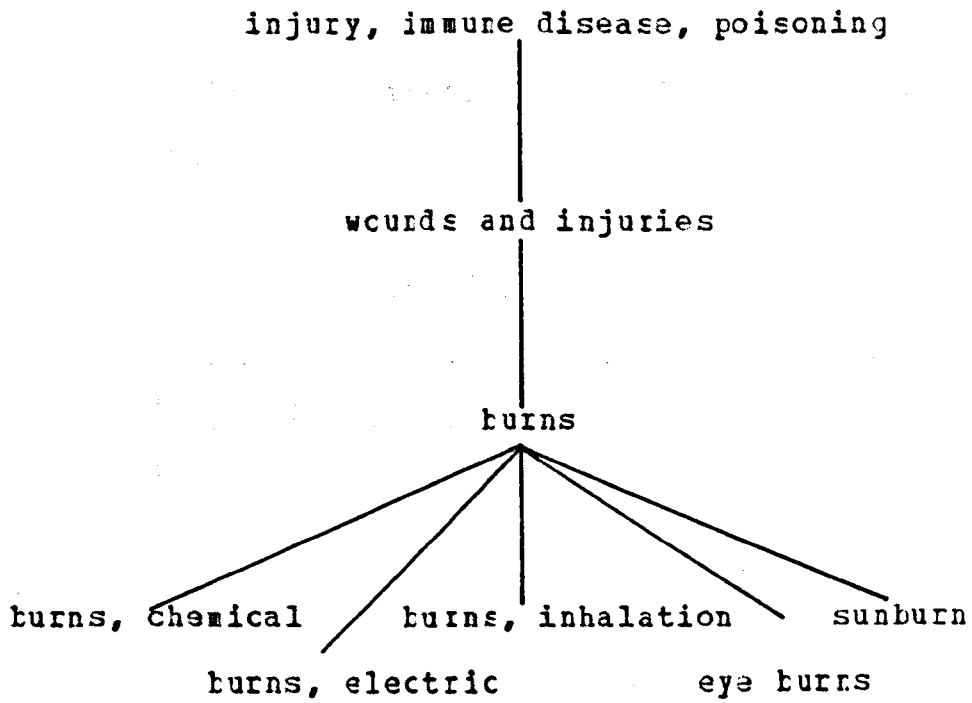


Figure 3.20.

In this case, "burns" would not be assigned to a document if the document was about a particular type of burn. A more specific term, e.g. "sunburn", would be used instead. It is unlikely that both "burns" and "sunburn" or indeed any of the four other terms at that level will occur together in the same document.

For a given collection size, the number of possible combinations is a function of the number of terms used (t)

in that collection.

$C = (t(t-1))/2$

For the sub-collections, the following figures are obtained (Table 3.2)

Combination statistics

no. of docs	no. of terms used	max. no. of possible combs.	actual no. of combs.	%age
500	2128	2263128	18463	0.82
5000	6521	21258460	184220	0.87
10000	7596	28845810	340129	1.18
15000	8129	33036256	483273	1.46
20000	8422	35460831	591924	1.67
25000	8639	37311841	705132	1.89

Table 3.2.

As the number of documents in the sub-collections increases, the number of different terms used increases,

but levels out (Figure 3.19). The maximum number of possible different combinations is a function of this figure and increases in proportion to the square of the number of terms. The actual number of different combinations in the sub-collections is but a small percentage of the maximum possible and shows no sign of levelling off.

3.6 PRACTICAL EXPERIMENTS ON TERM COMBINATIONS.

3.6.1 Experimentation note

Because of the high demands in both computation time and storage made by the experiments on term combinations, they were restricted to the sub-collections only. (There were over 700,000 different combinations in the 25000 sub-collection.)

3.6.2 "New combinations"

In 3.4.2 the introduction of new terms was investigated. In the same way, the introduction of

combinations was studied under the assumption that a collection, to which the addition of documents showed little increase in the number of different combinations, could be deemed representative of the full collection.

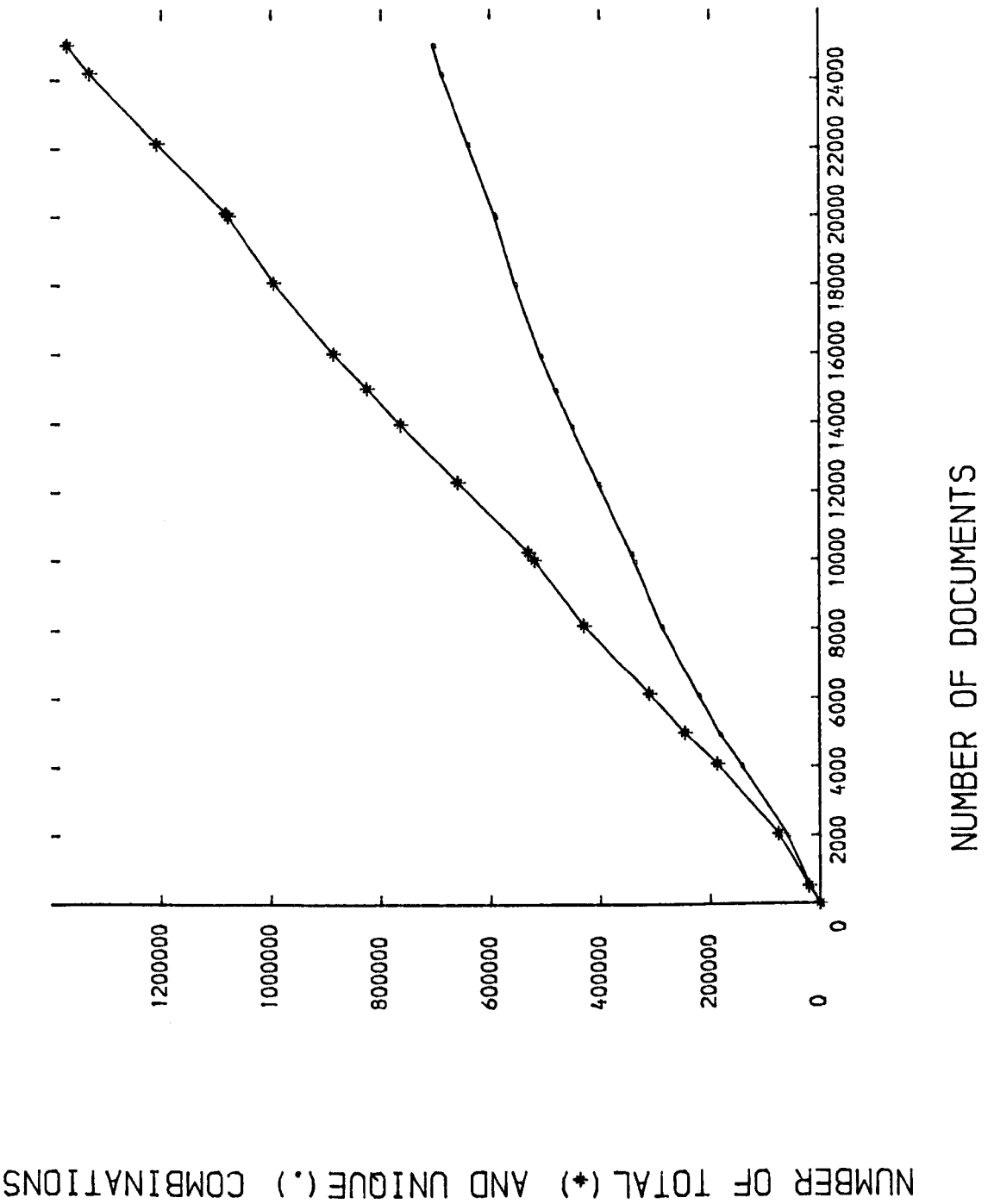
Table 3.3 gives details of both the number of different combinations used and the total number of combinations occurring in the sub-collections, which are shown graphically in Figure 3.21 (once again with intermediate points at file boundaries).

Number of documents	Number of different combinations	Total number of combinations	Ratio
500	18463	20074	91.98
5000	184220	246161	74.84
10000	340129	520727	65.32
15000	483273	825229	58.56
20000	591924	1078296	54.89
25000	705132	1370795	51.44

Table 3.3.

In marked contrast to the corresponding graph for single terms (Figure 3.19), there is no levelling out between 15000 and 20000 documents. Indeed, the graph

Figure 3.21


NUMBER OF TOTAL (*) AND UNIQUE (.) COMBINATIONS
0
200000
400000
600000
800000
1000000
1200000
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
NUMBER OF DOCUMENTS

remains almost linear as far as the 25000 limit, and any slight deviation from a straight line is matched by a similar deviation in the total number of combinations.

3.6.3 Combination frequencies

Because of the vast number of combinations involved (700,000+) and the amount of computing resources that would have been necessary, it was impossible to perform a frequency analysis on all the combinations, following the method used for single terms (3.4.1).

In an attempt to solve this problem, it was decided to examine the frequencies of selected combinations which were derived from genuine query formulations as submitted to the MEDUSA and MEDLINE systems. Only those terms which, upon expansion, were combined together using "AND" and would retrieve documents, were selected so that a query of the form

R1 = (M1 OR M2) AND M3

yielded the combinations M1 + M3 and M2 + M3. Note - M1 + M2 was not included because a document containing those two terms and not M3 would not be retrieved.

A total of 117 query formulations were processed in this way yielding 407 different combinations. However, only 137 of these appeared in the 25000 sub-collection.

In theory, once a collection begins to repeat itself, i.e. a representative subset has been found, the frequency of a particular combination should increase in proportion to the increase in the number of documents in the collection.

This proved not to be the case. Indeed, the vast majority of the 137 combinations exhibited anything but a regular rate of increase. Only one combination was regular in this respect and that occurred only once in each month.

To throw further light on this, it was decided to perform actual MEDUSA searches using the combinations "anded" together as query input. MEDUSA retrieves the documents for each month separately - a document retrieved corresponds to that combination it represents being present. This enabled the analysis to be extended beyond the previous 25000 limit with additional points at 28421, 45718 and 61036 documents, but no improvement was evident.

3.6.4 Check for randomness.

At this stage, it was appropriate to include a specific check to prove that the documents were indeed randomly distributed and not grouped in ary way, for example by subject area. Documents other than those written in English had been excluded (see 3.3.5). To do this, a histogram was drawn for each combination, showing the number of times it occurred each month (MEDUSA retrieves documents month by month so the production of this was trivial). If each month was a representative of the full collection, then the number of occurrences per month would be relatively constant As expected from the results of the previous experiment, this was not the case.

To check for randomness, the documents were assigned to four groups randomly using a function of the citation number, instead of on a monthly basis. No uniformity was apparent, the results remaining roughly the same. This may be due to the fact that the combinations occur so infrequently that they cannot be anything but random and as such can be taken as an indication that the documents are not grouped in any way.

3.6.5 Frequency Analysis.

A remarkable feature of the work concerning the combinations extracted from the queries is the low frequency with which they occur.

To investigate this, a frequency analysis was performed on the largest sub-collection for which figures were available, the 25000 document set. A summary of results is shown below.

Number of documents = 25000
Number of different combinations = 705132
Total number of combinations = 1370795

497931 combinations occur only once,
102170 combinations occur twice,
39532 combinations occur 3 times,
20002 combinations occur 4 times,...
1 combination occurs 440 times (max.)

approx. 70% of combinations occur once only,
approx. 98% of combinations occur less than 10 times.

This serves tc confirm the rejection of combinaticns of more than two terms as objects of enquiry.

3.6.6 Full query searches.

Searches were also performed on the monthly collections using complete queries rather than their constituent combinations. This provided a means of determining the relative importance of each combination (and therefore each term) within the query and investigating whether this varied from month to month.

As an example, consider the following:-

Q1=M38 AND(M2 OR M5 OR M7 OR M8 OR M9 OR M10 OR M11)

Combinations which retrieve documents:-

	Month1	Month2	Month3	Month4	Total
C1: M38 AND M2	1	0	0	0	1
C2: M38 AND M7	1	0	0	0	1
C3: M38 AND M9	0	1	1	2	4
C4: M38 AND M10	2	0	2	2	6
Total incl.	-	-	-	-	-
duplicate docs.	4	1	3	4	12
Q1 (full query)	3	1	2	3	9

Whilst the four combinations retrieve 4, 1, 3, 4 documents for each month respectively, only 3, 1, 2, 3 different documents are retrieved. This is shown in Figure 3.21.

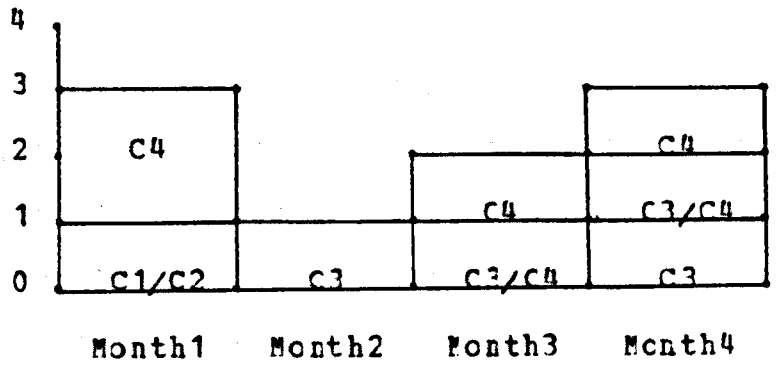


Figure 3.21.

In Month 1, C4 retrieves 2 documents, C1 and C2 retrieve the same document, a total of 3 documents.
In Month 2, C3 retrieves 1 document.
In Month 3, C4 retrieves 2 documents, one of which is also retrieved by C3.
In Month 4, C4 and C3 each retrieve 2 documents, one of which is common to both.

C4 is the most important combination in that it retrieves the most documents over three months only. It retrieves no documents in Month 2. None of the seven queries shows any consistency over the months as far as the most important combination is concerned. This further

confirms that a monthly section (15,000 documents) is not representative of the full collection.

3.7 SUMMARY OF RESULTS.

This chapter has attempted to give an indication of the relationships between theoretical measures and their empirically derived equivalents. In general, the actual values differ considerably from those that have been theoretically predicted. Nevertheless, the theoretical work has proved its worth by pointing the practical experiments in the right direction and giving an idea of what kind of behaviour to expect.

The practical experiments contained in this chapter have not produced any conclusive results as to the size of collection required in order to be able to confidently predict retrieval behaviour in an operational environment. The single term experiments revealed a possible cut-off point between 15,000 and 20,000 documents but this was not supported by the work on two-term combinations. The actual searches revealed not only the inconsistencies in the collection under growth but also the very low frequencies with which combinations occurred and the subsequently small number of documents retrieved by

genuine search formulations.

4 EXPERIMENTS USING DOCUMENT CLUSTERING

4.1 INTRODUCTION

Whilst the previous chapter was primarily concerned with the nature of collections in terms of the characteristics of the documents they contain, i.e. the document was the unit of study, this chapter attempts to investigate document collections by looking at their overall structure and the inter-document relationships within the collection and how these vary according to collection size.

To this end, a clustering method has been applied to several small sub-collections. Document clustering is used to reduce the amount of searching required in retrieval by dividing the collection into groups of related documents ("clusters"). This is equivalent to dividing the subject area into more specialised topics. The results of such experiments may be compared with results of other experiments in clustering.

If any sub-collection constitutes a true subset of the full collection and therefore may be used meaningfully for

experimental purposes, the structure of the collection will remain stable if further documents are incorporated. In terms of the effect of an increase in collection size on the clusters, the number of clusters should remain constant, but the size of each cluster should increase in proportion to the number of documents added to the collection. In other words, large collections should have the same number of clusters as small collections, but the number of documents they contain should be greater. This is equivalent to the non-introduction of new topics into a subject area, and an increase in the number of articles concerned with each topic. This is of course, the ideal situation - from a more practical viewpoint, minor changes in the structure may be apparent.

4.2 CHOICE OF CLUSTERING METHOD

There are two, often conflicting, criteria governing the selection of a clustering method. Firstly, it should exhibit a degree of theoretical soundness. In order to do this, it should satisfy the following criteria of adequacy, which have been adapted for use in IR by van Rijsbergen (1979) from those proposed by Jardine and Sibson (1971) in their discussion of general classification techniques:-

4. The method produces a clustering which is unlikely to be altered drastically by the incorporation of further documents, i.e. it is stable under growth.

5. The method is stable, i.e. small changes in the indexing of the documents lead to small changes in the clustering.

6. The method is independent of the initial ordering of the documents.

The second criterion is that the method should be efficient in terms of computation time and storage requirements. Unfortunately, this has proven to be the overriding consideration in many clustering methods, with the effect that there are many very efficient methods now available, which do not satisfy the requirement of theoretical soundness. These are termed heuristic clustering methods.

4.2.1 Heuristic cluster methods

Because of their pre-occupation with efficiency, heuristic cluster methods rarely have any theoretical background and are often defined in terms of the

algorithms implementing them. They proceed directly from the document descriptions to the clustering without any intermediate stage. They attempt to speed up the retrieval process by limiting the extent of a linear search.

In linear associative retrieval (LAR), each search request must be matched with each and every document in the collection, in order to determine the set of relevant documents. By clustering the collection, documents are assigned to groups of related documents, and for each cluster a cluster representative is derived. The role of the cluster representative is to summarise the content of the documents contained within that cluster. This can be achieved by using the centroid of the cluster (cf. centre of gravity) or the document with the most attributes in common with the rest. The search request is then matched with the set of cluster representatives and the set of relevant documents is ultimately selected by matching the search request with the members of those clusters, which gave the best match. This process reduces dramatically the number of comparisons between request and document but may lead to a degradation of retrieval effectiveness compared to LAR.

An important point to consider in the case of heuristic clustering algorithms is that they do not attempt to extract a structure from the document collection, rather they try to impose a suitable structure upon it. This may be achieved by prespecifying parameters, which have to be empirically determined. These include limits on the number of clusters, the maximum and minimum size of clusters, the threshold value of the matching function for a document to be included in a cluster and the degree of overlap between clusters.

4.2.2 Matching Functions.

Before proceeding, it is important to explain the concept of a matching function (similarity coefficient, association measure). A matching function measures the association between a document and either a cluster representative (as in this case), or another document. Several different matching functions have been proposed for use in IR, but providing they are correctly normalised with respect to the number of terms in the documents, there is little effect on retrieval performance to be gained by using one instead of another. Thus, the similarity between two documents or between a document and a cluster representative, X and Y, may be expressed in a

number of ways:

$|X \cap Y|$ - Simple matching function

$\frac{2.|X \cap Y|}{|X|+|Y|}$ - Dice's coefficient

$\frac{|X \cap Y|}{|X \cup Y|}$ - Jaccard's coefficient

$\frac{|X \cap Y|}{|X|^{\frac{1}{2}}.\ |Y|^{\frac{1}{2}}}$ - Cosine coefficient

$\frac{|X \cap Y|}{\min(|X|,|Y|)}$ - Overlap coefficient

(X,Y = set of keywords representing documents / cluster representatives,

$|.|$ = counting measure)

The second and subsequent measures can be seen as normalised versions of the simple matching function.

4.2.3

The action of a heuristic clustering algorithm is best explained by an example. Rocchio's algorithm (Rocchio 1966) is one of the best known. It has three stages :-

1. A number of documents are selected (by some criterion) as cluster representatives. The remaining documents are assigned either to existing clusters or to a "rag bag" cluster for misfits, by thresholding a matching function. A document may be assigned to more than one cluster.

2. The resulting clustering is adjusted to comply with the prior specification parameters.

3. The clustering is tidied up by forcibly assigning documents from the "rag bag" cluster and reducing the overlap between clusters.

Much of the research in the area of heuristic algorithms has been towards reducing the number of passes necessary to produce a clustering, and several "single pass" algorithms have been proposed.

4.2.4 Disadvantages of Heuristic Algorithms

Whilst they undoubtedly produce a clustering very quickly and efficiently, heuristic algorithms have the following disadvantages, which makes their use in this research undesirable :-

1. The clustering is dependent upon the order in which the documents are submitted to the algorithm.

2. They are not stable under growth. Updating of a heuristic clustering to incorporate new documents is difficult. Often a complete reclassification is necessary to correct the clustering after one or more updates.

3. The clustering is dependent upon the specification of parameters, which need to be known in advance and may vary from collection to collection.

4. A structure is imposed upon the collection according to these parameters.

These factors combine to place in doubt whether changes in structure are due to differences in collection size or

inconsistencies in the clustering algorithm.

4.3 HIERARCHIC CLUSTERING METHODS.

There are many hierarchic cluster methods but by far the most popular and most documented in IR is the single-link method. This has been shown to satisfy all the criteria of adequacy for theoretical soundness (Jardine and Sibson, 1971). Rather than having a detrimental effect on retrieval performance in terms of recall and precision, Jardine and van Rijsbergen (1971) stated that such a method had the potential for improving effectiveness in comparison with a linear search. To reinforce their argument they postulated the Cluster Hypothesis, which states, "Closely associated documents tend to be relevant to the same requests". This explcits the separation within a collection between relevant and non-relevant documents (van Rijsbergen and Sparck Jones, 1973).

4.3.1 Single-link.

Instead of operating on a measure of similarity between objects, single-link takes as its input a matrix of

dissimilarity coefficients (IC's) showing the degree of dissimilarity between documents. The matrix contains n(n-1)/2 elements for a collection of n documents. It is the generation of this matrix that disadvantages hierarchic methods in general, with regard to their computational efficiency compared to heuristic algorithms.

From this matrix, a hierarchy of non-overlapping clusters is produced with each level having an associated numerical value. This value gives a measure of the association between the documents contained in the clusters at that level. The hierarchy may be represented geometrically by a dendrogram. This is easily translated into a tree structure, which makes this method particularly attractive for use in operational systems, because of the efficient search stategies that have been devised for trees.

The action of the single-link method is best described by means of an example:

Consider a collection of 5 documents (A,B,C,D,E), analysis of which produces the following dissimilarity matrix (Figure 4.1) :

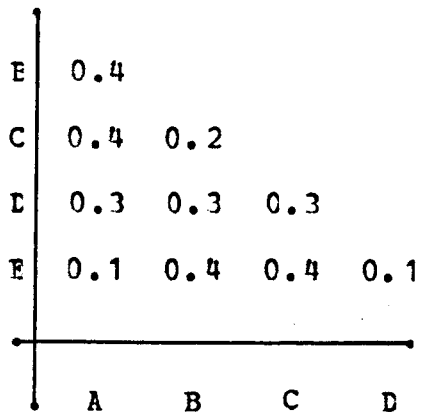


Figure 4.1.

By thresholding the DC's at levels of 0.1, 0.2, and 0.3 graphs can be drawn (Figures 4.2 to 4.4)

0.1

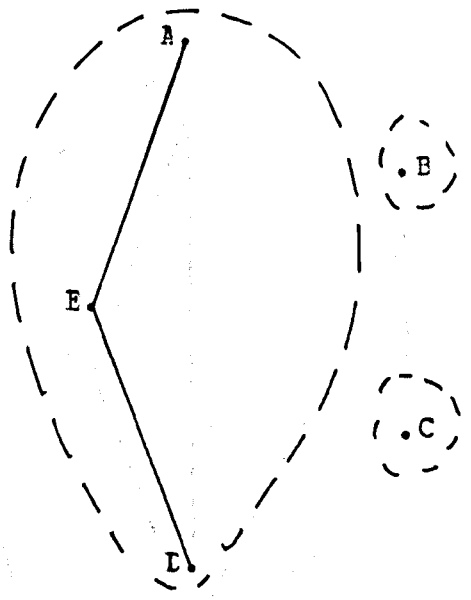


Figure 4.2.

0.2

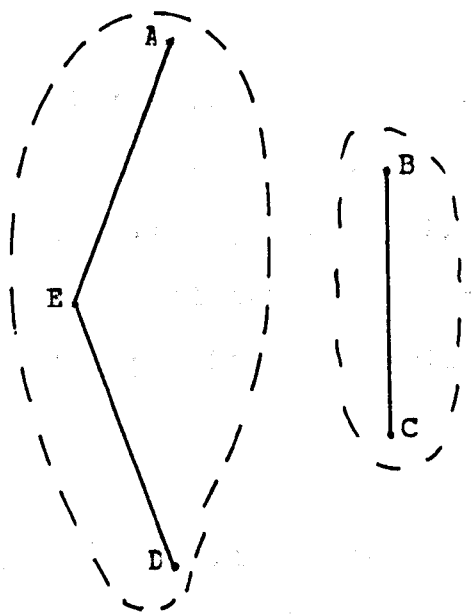


Figure 4.3.

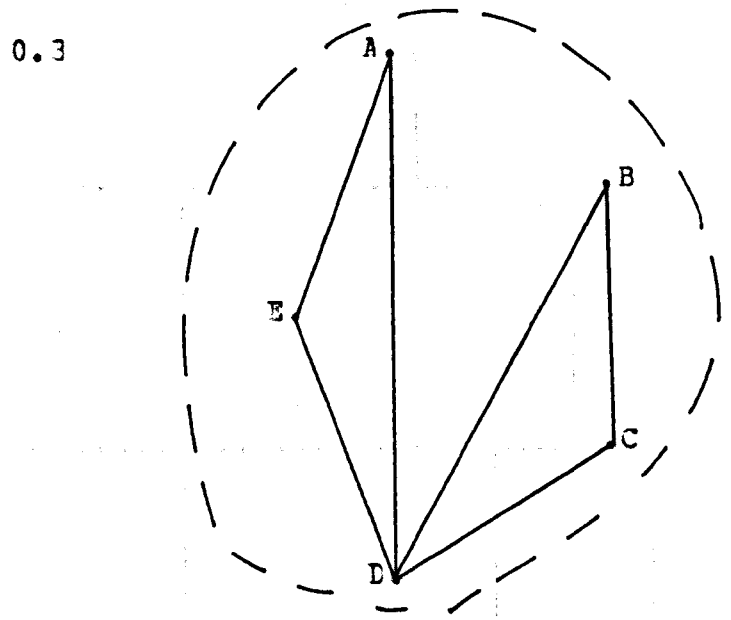


Figure 4.4.

The resulting clusters are surrounded by dotted lines. From this it can be seen that for a document to be included in a cluster at a given level, it need only be associated with one member of the cluster with a DC less than or equal to the particular level, hence, the term "single-link".

The graphs may be converted into a dendrogram, which enables the overall structure to be examined (Figure 4.5).

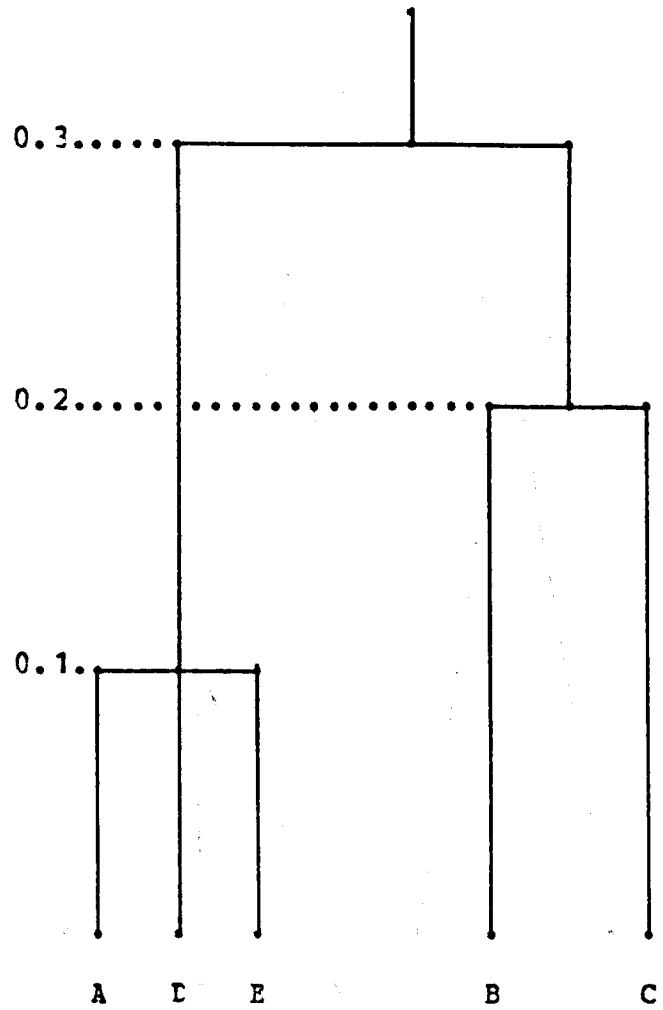


Figure 4.5.

This can be translated into a suitable tree structure (Figure 4.6).

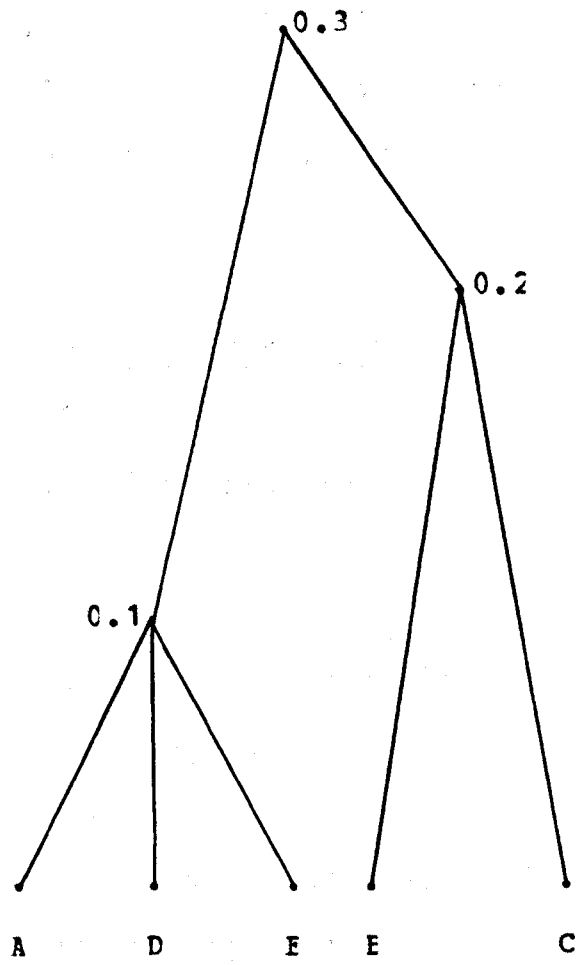


Figure 4.6.

Note that it is not necessary to threshold the matrix at 0.4 as this does not alter the clustering of 0.3, where all the documents are included in one cluster.

4.3.2

The choice of IC has little effect on retrieval performance, and indeed, single-link ensures that the clustering does not depend on the actual values of the IC's but upon their rank-ordering.

4.4 APPROPRIATENESS OF SINGLE-LINK FOR THIS RESEARCH.

Single-link clustering was chosen for the following reasons:

1. It is the only hierarchic method to satisfy the criteria of adequacy.

2. It seeks a structure from the collection rather than attempting to impose one. An imposed structure could mean that results may be due to the clustering method, rather than differences in collection size.

3. It is order-independent.

4. It is stable under growth. This ensures that any variations are due to differences in collection size and not as a result of inconsistencies in the updating

mechanism of the method.

4.5 IMPLEMENTATION OF THE SINGLE-LINK METHOD.

The generation of the dissimilarity matrix was performed by using Willett's algorithm (Willett, 1981). By utilising the inverted file, the number of matches is reduced considerably by omitting the calculation of DC's between documents with no terms in common.

The choice of DC was made purely on the grounds of computational convenience. The following DC was used:

$$DC = \frac{|X \Delta Y|}{|X|+|Y|}$$

where $|X \Delta Y| = |X \cup Y| - |X \cap Y|$

It is related to Dice's coefficient by

$$\frac{|X \Delta Y|}{|X|+|Y|} = 1 - \frac{2.|X \cap Y|}{|X| + |Y|}$$

The single-link method was implemented by using the SLINK algorithm due to Sibson (1973). Although this algorithm must be supplied with DC's in a specific order, it does discard them as soon as they are processed, thus reducing storage requirements. Furthermore, the output is readily translated into a dendrogam representation, which allows a visual examination of the clustering to be made. As no searching of the collections was necessary, an elaborate tree structure implementation was not required. For this reason, van Rijsbergen's DYNALINK algorithm (van Rijsbergen, 1971), which otherwise may have been chosen for its ability to accept DC's in any order, was rejected.

4.6 THE EFFECTS OF SIZE ON THE CLUSTER STRUCTURE OF COLLECTIONS.

As a result of the comparative inefficiency of the single-link method, substantial overheads of computation time restricted the experiments involving document clustering to sub-collections containing 250, 500, 1000 and 5000 documents.

As an example of this, the generation of the dissimilarity matrix for the 5000 sub-collection took 618 CPU seconds, and the single-link clustering a further 1125

seconds, using the IBM 370/168 running under MTS at Newcastle. Much of the time was taken up by I/O operations and it is felt that a substantial improvement could be obtained if the two tasks were combined into the same program.

The results of Chapter 3 indicate that unchanging clusters are unlikely to occur unless very large collections are clustered, certainly larger than 5,000 documents. However, much of the work on clustering has been performed on collections containing relatively small numbers of documents, and therefore collections of this size are worth considering.

Figure 4.7 gives an example of a dendrogram, for the 250 sub-collection. Visual inspection reveals large differences between the clustering structures. The 250 sub-collection has very few clusters at lower levels, the number of which increases with collection size. Perhaps, the most readily perceivable charcteristic of all the structures is the lack of large clusters until quite high levels of dissimilarity (0.7 upwards).

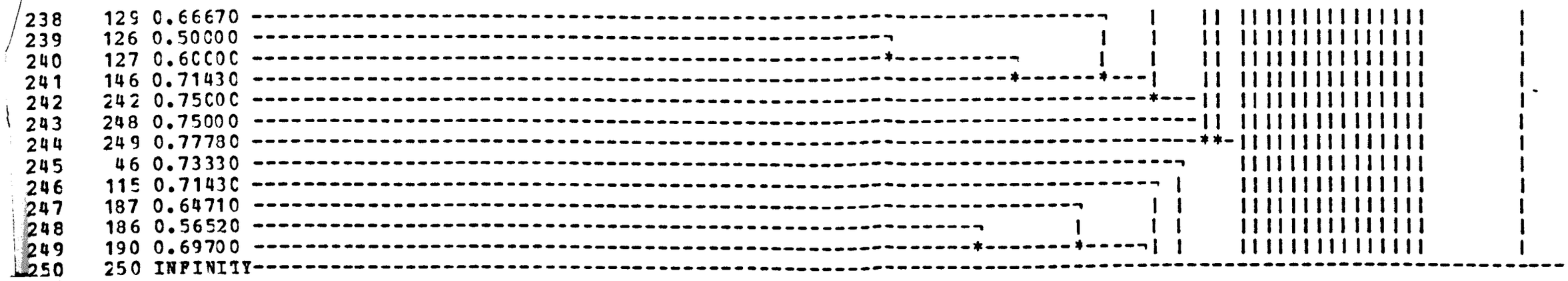
PULLOUT

ne	Dcc#	level	0123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
1	44	1.00000	
2	51	1.00000	
3	78	1.00000	
4	159	1.00000	
5	172	1.00000	
6	218	1.00000	
7	241	1.00000	
8	29	0.88240	
9	189	0.91670	
10	206	0.90910	
11	86	0.90480	
12	12	0.90000	
13	28	0.90000	
14	98	0.90000	
15	182	0.90000	
16	15	0.89470	
17	22	0.89470	
18	4	0.88890	
19	24	0.88890	
20	10	0.83330	
21	14	0.86670	
22	141	0.81820	
23	216	0.88890	
24	233	0.87500	
25	2	0.83330	
26	245	0.87100	
27	45	0.86670	
28	62	0.86670	
29	63	0.86670	
30	183	0.85710	
31	184	0.86670	
32	192	0.86670	
33	213	0.86670	
34	27	0.86210	
35	100	0.85710	
36	47	0.77780	
37	111	0.85710	
38	191	0.85710	
39	204	0.85710	
40	238	0.85710	
41	17	0.85190	
42	79	0.66670	
43	89	0.85190	
44	225	0.85190	
45	20	0.84620	
46	101	0.84620	
47	119	0.84620	
48	69	0.80000	
49	102	0.83330	
50	30	0.80950	
51	121	0.84620	
52	147	0.84620	
53	43	0.75000	
54	171	0.84620	
55	198	0.84620	
56	215	0.84620	
57	105	0.82610	

58	244	0.84620	
59	31	0.72220	
60	32	0.82860	
61	90	0.84210	
62	3	0.76920	
63	18	0.84000	
64	169	0.84000	
65	1	0.83330	
66	7	0.20000	
67	9	0.83330	
68	40	0.83330	
69	70	0.83330	
70	85	0.83330	
71	113	0.69230	
72	134	0.83330	
73	138	0.83330	
74	153	0.83330	
75	193	0.83330	
76	114	0.81820	
77	202	0.83330	
78	5	0.80000	
79	237	0.83330	
80	247	0.83330	
81	230	0.82860	
82	16	0.82610	
83	120	0.82610	
84	123	0.82610	
85	125	0.82610	
86	188	0.82610	
87	243	0.82610	
88	35	0.81820	
89	68	0.81820	
90	73	0.81820	
91	103	0.81820	
92	118	0.81820	
93	149	0.81820	
94	41	0.72410	
95	161	0.81820	
96	173	0.81820	
97	177	0.81820	
98	185	0.81820	
99	110	0.78950	
100	19	0.75000	
101	152	0.70000	
102	88	0.64710	
103	219	0.81820	
104	222	0.81820	
105	49	0.80950	
106	150	0.80950	
107	199	0.80950	
108	205	0.80950	
109	112	0.80650	
110	25	0.80000	
111	42	0.80000	
112	53	0.77780	
113	54	0.80000	
114	76	0.80000	
115	164	0.80000	
116	194	0.80000	
117	200	0.80000	

118	201	0.80000	
119	210	0.80000	
120	221	0.80000	
121	226	0.80000	
122	227	0.80000	
123	94	0.79310	
124	75	0.73330	
125	81	0.78950	
126	82	0.78950	
127	87	0.78950	
128	95	0.78950	
129	99	0.78950	
130	231	0.78950	
131	74	0.77780	
132	65	0.71430	
133	130	0.77780	
134	13	0.66670	
135	26	0.66670	
136	80	0.66670	
137	145	0.77780	
138	168	0.77780	
139	170	0.77780	
140	208	0.77780	
141	223	0.77780	
142	232	0.77780	
143	55	0.73910	
144	246	0.77780	
145	57	0.61900	
146	144	0.76470	
147	148	0.76470	
148	61	0.52380	
149	116	0.66670	
150	151	0.76470	
151	8	0.75000	
152	36	0.75000	
153	92	0.75000	
154	38	0.33330	
155	39	0.42860	
156	93	0.75000	
157	106	0.75000	
158	108	0.75000	
159	117	0.75000	
160	128	0.75000	
161	157	0.75000	
162	21	0.55560	
163	162	0.75000	
164	165	0.75000	
165	209	0.75000	
166	224	0.75000	
167	228	0.54550	
168	229	0.75000	
169	104	0.73910	
170	235	0.75000	
171	236	0.75000	
172	56	0.64290	
173	167	0.73910	
174	97	0.73330	
175	107	0.73330	
176	109	0.73330	
177	60	0.71430	

178	133	0.73330
179	6	0.63640
180	83	0.71430
181	77	0.69230
182	154	0.71430
183	176	0.68420
184	143	0.66670
185	142	0.64710
186	195	0.71430
187	50	0.66670
188	64	0.60000
189	91	0.69230
190	214	0.71430
191	155	0.64710
192	239	0.73330
193	33	0.71430
194	37	0.71430
195	72	0.71430
196	96	0.71430
197	122	0.71430
198	124	0.71430
199	137	0.71430
200	156	0.71430
201	158	0.71430
202	34	0.69230
203	211	0.70000
204	71	0.68420
205	212	0.71430
206	174	0.57890
207	175	0.64710
208	179	0.60000
209	180	0.66670
210	181	0.66670
211	196	0.66670
212	48	0.60000
213	203	0.66670
214	52	0.65220
215	217	0.71430
216	132	0.70000
217	160	0.70000
218	207	0.70000
219	197	0.69230
220	135	0.66670
221	139	0.66670
222	140	0.66670
223	11	0.61900
224	23	0.57890
225	58	0.62500
226	84	0.64710
227	163	0.62500
228	166	0.66670
229	136	0.63640
230	220	0.69230
231	59	0.66670
232	66	0.55560
233	67	0.66670
234	178	0.68420
235	234	0.71430
236	240	0.75000
237	131	0.71430


238 129 0.66670
239 126 0.50000
240 127 0.60000
241 146 0.71430
242 242 0.75000
243 248 0.75000
244 249 0.77780
245 46 0.73330
246 115 0.71430
247 187 0.64710
248 186 0.56520
249 190 0.69700
250 250 INFINITY

A more objective method of analysing the structures is possible by considering the following statistics for each sub-collection:

1. The percentage of documents clustered at a given level. This measure is commonly used in clustering experiments and is calculated by determining the level at which each document becomes connected into the hierarchy.

2. The percentage reduction in the number of branches in the dendrogram/tree structure. Given that there are as many branches as documents at the lowest level, the number of branches remaining at a given level can be calculated, and hence, the reduction. The number of branches is made up of the number of clusters plus the number of unclustered documents.

3. The number of clusters at a given level.

4. The number of documents contained in the clusters at a given level.

Having introduced these measures, it is worthwhile re-stating the characteristics of a true collection subset

from the point of view of single-link clustering. At a given level in the hierarchy, the point at which the number of clusters becomes constant despite the addition of further documents to the collection, and where only the clusters themselves increase in size, can be deemed a true subset and therefore representative for experimental purposes. Once a subset reaches a size which is representative, additional documents will be incorporated into the hierarchy at lower levels. Until this point is reached, however, an increase in the percentage of documents clustered at lower levels should be apparent as more and more clusters are formed. This should result in a corresponding increase in the percentage reduction in branches.

1) and 2) above are normalised measures, which allow comparison between collections of different sizes. As a further aid to comparison, arbitrary levels of DC (0.1, 0.2 ..... 0.9) have been selected as points at which to extract the following statistics (Tables 4.1 to 4.4) :-

Collection: MEDLARS250

Number of documents: 250

DC level	%age clstrd	%age rdctn	nc. clstrs	size range of clusters
0.1	.	.	.	.
0.2	0.8	0.4	1	1 cluster 2 docs
0.3	0.8	0.4	1	1*2
0.4	1.6	0.8	2	2*2
0.5	2.8	1.6	3	2*2, 1*3
0.6	11.2	6.0	13	11*2, 2*3
0.7	33.2	20.8	21	10*2, 1*20
0.8	66.4	61.6	12	10*2, 1*141
0.9	96.4	95.5	3	1*2. 1*239

Table 4.1.

Collection: MEDLARS500

Number of documents: 500

DC level	%age clstrd	%age rdctn	no. clstrs	size range of clusters
0.1	0.4	0.2	1	1*2
0.2	0.8	0.4	2	2*2
0.3	1.2	0.6	3	3*2
0.4	3.2	1.6	8	8*2
0.5	6.8	3.6	16	14*2, 2*3
0.6	18.4	11.0	37	26*2, 2*5
0.7	47.6	40.0	38	22*2, 1*131
0.8	81.6	79.4	11	10*2, 1*388
0.9	99.0	98.8	1	1*495

Table 4.2.

Collection: MEDLARS1000

Number of documents: 1000

DC level	%age clstrd	%age rdctn	no. clstrs	size range of clusters
0.1	0.2	0.1	1	1*2
0.2	1.0	0.5	5	5*2
0.3	1.8	0.9	9	9*2
0.4	5.3	2.7	26	25*2, 1*3
0.5	13.9	8.2	57	47*2, 1*8
0.6	33.7	24.5	92	66*2, 1*83
0.7	67.3	61.3	60	40*2, 1*503
0.8	93.3	92.5	8	7*2, 1*919
0.9	99.7	99.6	1	1*997

Table 4.3.

Collection: MEDLARS5000

Number of documents: 5000

DC level	%age clstrd	%age rdctn	no. clstrs	size range of clusters
0.1	0.9	0.5	21	20*2, 1*3
0.2	1.8	1.0	41	34*2, 1*5
0.3	3.3	1.8	74	63*2, 1*5
0.4	8.0	4.8	159	120*2, 1*10
0.5	20.4	13.8	328	237*2, 1*37
0.6	44.5	36.6	393	266*2, 1*1087
0.7	78.9	76.1	138	103*2, 1*3602
0.8	99.0	98.5	1	1*4928
0.9	100.0	99.9	1	1*4999

Table 4.4.

A feature of the clustering structures is that as the collection size increases, more and more documents become linked into the hierarchy at lower levels and more clusters are formed. In the case of a true subset, one would expect nearly all additional documents to be incorporated at these low levels, as the collection begins to repeat itself. In the case of the four sub-collections analysed here, the general trend is certainly towards more clusters at lower levels, but there is no sign of the figures reaching a constant value, which would be the case if a true subset had been attained. This would appear to indicate that sub-collections containing less than 5000 MEDLARS documents are too small to be representative. This is perhaps to be expected in the light of the experiments on single terms and 2-term combinations and the very small size of the sub-collections used in these clustering experiments.

It is also interesting to examine the sizes of the clusters. At low levels (up to C.5), there is a large number of very small clusters, typically containing only two documents, over all the sub-collections. As expected, the larger collections generally have more clusters at a particular level, but even then it is only at the 0.6 level in the 1000 sub-collection and at the 0.5 level in the 5000 sub-collection that what may be termed reasonably

sized clusters are formed. However, at these levels only 33.7% and 20.4% respectively of documents are clustered. At higher levels, all the clusters are amalgamated into one large cluster containing upwards of 20% of the collection. Since the retrieval mechanism in systems using hierarchic clustering is to retrieve all the documents in the chosen cluster, clusters of (say) 5 to 10 documents would be of most use. With the clustering produced here, it would be common to retrieve either clusters with very few documents in them, or a cluster containing most of the documents in the collection.

4.6.1 Comparisons with other experiments.

There have been numerous experimerts in the area of document clustering using single-link, chiefly at Cambridge. Whilst most of these have been concerned with the retrieval effectiveness of single-link clustering and as such, their results tend to be expressed in these terms, figures for the percentage of documents clustered and in one case, the percentage reduction in branches, are available. These are reproduced in Tables 4.5 and 4.6.

Percentage of Documents Clustered

Lvl	Cranfield 200 Docs 32 t/d (a)	INSPEC 541 Docs 12.2 t/d (a)	Keen 797 Docs 7.2 t/d (a)	Cranfield 1400 Docs 53.6 t/d (b)	UKCIS 11613 Docs (c)
0.1	.	.	.	0.57	.
0.2	3.0	.	.	1.29	2.0
0.3	6.5	.	4.0	3.71	4.0
0.4	13.5	3.0	17.5	11.93	12.0
0.7	40.0	8.0	35.5	40.86	33.0
0.6	74.0	20.5	74.5	77.93	65.0
0.7	93.5	63.0	92.5	96.43	91.0
0.8	99.5	98.5	100.0	99.54	N/A
0.9	100.0	100.0	100.0	100.00	N/A

Sources

a = Sparck Jones (1973)

b = van Rijsbergen & Croft (1975)

c = Croft (1977)

Table 4.5.

Sparck Jones (1973) also gives the percentage reduction in branches.

Percentage reduction in branches.

Lvl	Cranfield 200	INSPEC 541	Keen 797
0.1	.	.	.
0.2	1.5	.	1.8
0.3	3.5	.	6.3
0.4	8.5	2.4	19.1
0.5	27.0	8.5	47.4
0.6	65.5	29.2	82.9
0.7	92.0	66.5	97.6
0.8	99.0	98.8	100.0
0.9	100.0	100.0	100.0

Table 4.6.

The results obtained for MEDLARS collections bear most resemblance to the figures for the INSPEC test collection.

This may be due to a similarity in the average number of terms per documents (INSPEC has 12.2 terms/document, MEDLARS between 9 and 10) or to the fact that both have been extracted from a large operational database.

The Cranfield figures are in general much higher than both the other two test collections and MEDLARS. This is possibly caused by the unusually high number of terms assigned to the documents. This increases the likelihood of documents having terms in common, and as a result, the degree of clustering.

Unfortunately, there are no figures available for the number of clusters and the number of documents each cluster contains. However, given the percentage of documents clustered and the percentage reduction in branches at a given level, the number of clusters at that level may be calculated. The number of clusters is obtained by subtracting the number of unclusterd documents from the number of branches remaining. There are apparently inconsistencies in the published figures, as the number of clusters for the INSPEC and Keen collections according to this calculation, are negative. The matter is presently under discussion with the author. The figures for the Cranfield 200 collection can be utilised and the number of clusters at each level is given in Table

4.7.

Number of clusters for the Cranfield 200 collection.

lvl	%age docs clstrd	no. docs clstrd (A)	no. docs NOT clstrd (B)	%age rductn in no. brnchs	rductn in no. brnchs (C)	no. brnchs left (D)	no. of clstrs (D-B)
0.1	.	.	.	.	.	.	.
0.2	3.0	6	194	1.5	3	197	3
0.3	6.5	13	187	3.5	7	193	6
0.4	13.5	27	173	8.5	17	183	10
0.5	40.0	80	120	27.0	54	146	26
0.6	74.0	148	52	65.5	131	69	17
0.7	93.5	187	13	92.0	184	16	3
0.8	99.5	199	1	99.0	198	2	1
0.9	100.0	200	0	100.0	199	1	1

Table 4.7.

The number of documents in each cluster cannot be derived from these figures. All that can be said is that, for

example, at level 0.6 there are 17 clusters made up of a total of 148 documents. If Cranfield follows MEDLARS, then there may well be one cluster with (say) 110 documents in it and 16 others with only 2 or 3 documents in them. This may or may not be the case.

4.7 CONCLUSIONS ON CLUSTERING.

As might have been expected from the findings of earlier experiments on single terms and 2-term combinations, there is no evidence to suggest that sub-collections containing up to 5000 documents show any sign of reaching a fixed clustering structure. This indicates that such collections are not representative of the full collection and as such are unlikely to enable confident conclusions to be made from experiments using them.

Due to the large amounts of computer resources used in these clustering experiments, sub-collections of more than 5000 documents had to be excluded from this research. Ideally, experiments should be performed on much larger collections, preferably containing up to 25000 documents.

Despite any shortcomings incurred by the restricted nature of the experiments, they have proved to be useful, if only in establishing guidelines for further research into the effects of collection size on document clustering. Certainly, it appears that the single-link method is the method to use in such research, because of its inherent stability.

5 CONCLUSIONS.

5.1 INTRODUCTION.

The purpose of this chapter is to draw together the results of the experiments described in Chapters 3 and 4 and examine the effect of this research on other research in the field of Information Retrieval.

In Chapter 2, the vast differences in size between small test collections used in IR experiments on the one hand and large commercial databases on the other, were highlighted, along with the problems associated with the use of test collections. The effects of size on a document collection (MEDIARS) have been examined in an attempt to determine if a minimum collection subset can be found which exhibits similar retrieval characteristics to the 'full' collection. Such a subset could be used in experiments to enable the performance of the system using the full collection to be predicted.

5.2 THE USE OF MEDIARS DOCUMENTS

A factor which needs to be taken into account when evaluating the results of the experiments is the degree to which MEDIARS documents are suitable for this type of research. The fact that MEDIARS is indexed manually using a strictly controlled thesaurus of MeSH terms is an advantage rather than a disadvantage, because the number of terms which may be used to index a collection is known (in this case, it is 10,137). This allows a cut-off point, at which all the terms, or more realistically, a large percentage of terms have been introduced, to be sought. This point may provide an indication of the size of collection subset necessary to be representative of the full collection. This is certainly not the case if a free indexing scheme is used to index the collection, as the dictionary size is not known in advance.

The experiments concerning the introduction of single terms and the single term frequencies rely heavily on the concept of the full collection. The full collection in this case contains 61,036 documents. It is felt that this is sufficiently large in itself and contains sufficiently more documents than the largest sub-collection (25,000 documents) to ensure that comparisons between it and any sub-collections are legitimate.

5.3 DISCUSSION OF RESULTS.

The experiments on the effects of size on the MEDLARS document collection reported in this thesis can be divided into three main areas of investigation :-

1. The behaviour of single terms.

2. The behaviour of 2-term combinations.

3. The clustering structure.

Before considering the results of these experiments individually, it is necessary to examine how they are interrelated.

The term is the basic building block of a collection. Each document is represented by a number of different terms and it is the variety of terms that gives each document its individual identity and more importantly from the IR point of view, distinguishes it from other documents in the collection. The way in which terms co-occur in documents is investigated in the experiments on 2-term combinations, and whereas the single term can be seen as the lowest level of analysis, the combination in

its capacity to represent the content of a document, can be seen as the next level, the document level. The highest level of analysis is the collection itself, which can be investigated in terms of its behaviour when clustered.

5.3.1 Single terms.

The experiments concerned with the behaviour of single terms enabled an examination of the effects of size to be made at the lowest level. They indicated that a collection containing between 15,000 and 20,000 documents was beginning to show similarities to the full collection. At this point, the rate of introduction of new terms was diminishing sufficiently to be able to recognise a cut-off point around 15,000 documents, where 80% of the terms in the dictionary were present in the collection. Similarly, the comparison of term frequencies showed a tendency for the graphs to begin to stabilise at around the same point.

The analysis of single terms is appropriate to methods using matching functions for searching, in particular weighted retrieval. In weighted retrieval, the index terms of all the documents in the collection are assigned weights in accordance with their ability to discriminate

one document from another. Thus instead of a document having a purely binary representation where terms are either present or absent ('0' or '1' in a vector representation) and each term is deemed to be equally important in characterising the content of the document, terms are given weights, which enables an ordering of terms to be established ranging from the most to the least discriminating.

The following weighting scheme has been shown to improve retrieval performance (van Rijsbergen (1979)):-

$$W = \log (N / n) + 1$$

where N is the number of documents in the collection
and n is the number of times the term occurs in the collection.

Because of the large fluctuations in the relative frequencies of the single terms in the 500 sub-collection compared with the full collection and even some of the larger sub-collections, the weighting of terms according to this scheme may be adversely affected. For example, given that the same document is present in both the 500 sub-collection and the full collection (as is the case here), whilst it is probable that the terms will have been assigned different weight values because of the overriding

N factor in the weighting function, it is also possible that the actual order of the term weights is altered. Thus, the most discriminating term in the document when it was weighted as part of the 500 sub-collection may not necessarily remain so in the full collection.

Index term weighting is usually incorporated into systems where searching is performed by matching function, which is able to process weighted terms. A ranking of documents in decreasing order of similarity to the search request is thresholded and documents above the cut-off point are retrieved. Recall and precision are determined by the number of relevant documents retrieved, but because of the inconsistencies of the term frequencies over differing collections the recall/precision figures may be misleading as in the following example:-

Suppose the following ranking is obtained from the 500 collection

A B C D E F*G H I .....

It is possible that the ordering of these documents may be different in a search of the full collection,

A X X X B X C X X D G X*X F X E X X I X H .....

(X denotes a document not included in the 500 sub-collection, and * the cut-off point).

Depending on the choice of threshold, different documents may be included. In the full collection, documents E and F are not retrieved but G is, whereas in the 500 sub-collection, E and F are retrieved and G is not. If G is relevant and E and F are not relevant, then the recall/precision figures are adversely affected in the 500 test.

5.3.2 2-term combinations.

As might have been expected from the results of the single term experiments (the number of combinations is O(n^2), the number of single terms is O(n), n is the dictionary size), the number of different combinations continued to increase almost linearly with collection size as far as the limit of the experiment at 25,000 documents. Certainly there was no apparent cut-off point. This happened despite the fact that all the possible combinations cannot occur not only because of the MEDLARS specificity rule which states that the most specific term should be assigned, but also because of the unlikelihood of a document relating to two vastly different subjects and the subsequent exclusion of the combinations of terms representing them e.g. experimental animals and parts of the body.

The experiments in which searches were performed to find the frequencies of certain combinations, revealed no stability and confirmed the findings of the other experiments on term combinations. It is important to mention here the very low frequencies with which the majority of combinations occurred (3.6.5). Because of this, experiments which are dependent on comparing numbers of documents containing particular combinations of terms or particular groups of combinations are open to criticism.

The study of combinations is appropriate to Boolean searching methods. Indeed, it is partly because of this that actual searches were performed. Traditionally, retrieval performance is measured in terms of recall and precision. To reiterate, recall is the proportion of relevant documents that are retrieved and precision is the proportion of retrieved documents that are relevant. A feature of the searches on MEDIARS collections is the very small number of documents that are actually retrieved. From a total of 117 initial searches, only 28 retrieved documents from the 25000 sub-collection. When the same searches were performed on the full collection of over 61,000 documents a total of 490 documents were retrieved, of which 222 were retrieved by one particular search statement. Of course, there is no relevance information

available to calculate recall and precision values for these searches, but because of the small number of documents retrieved, it is unlikely that any reasonable figures could be calculated.

5.3.3 Clustering structures.

The clustering structure is a representation of the collection as a whole and not just the individual documents and as such offers perhaps the best method of studying the document collection under growth. Unfortunately, because of excessive resource requirements, experiments using clustering were restricted to collections of up to 5,000 documents and the only conclusion that can be drawn as to the size of subset necessary for experiment, is that 5,000 documents is not a large enough collection. Nevertheless, if a clustering method could be applied to larger document collections, then it could well provide the best indication of the collection being a subset and therefore representative.

The behaviour of the clustering structures is related to the results of the combination frequency experiments. The dissimilarity coefficient is a function of the number of terms the two documents have in common. The abundance

of low frequency combinations (3.6.5) indicates that relatively few documents have more than a single term in common. This goes a long way to explain the small number of small clusters in the hierarchy until the level was reached where one large cluster was formed presumably at which point documents with only one term in common were starting to be incorporated in the clustering (see Tables 4.1 to 4.4).

This may also explain the differences in the Cranfield clustering. The Cranfield documents contain many more index terms (around 30) and the total number of different terms is considerably less than for MEDLARS. These combine to increase the probability that documents will have more than one term in common. It follows that documents may indeed have many terms in common and therefore will be incorporated into the clustering hierarchy at much lower levels than MEDLARS.

The most striking feature of the clustering of these admittedly small collections is the size of the clusters. At low levels in the hierachies, if a document is clustered at all it is only with one or possibly two others. This trend continues up the hierarchy until the point is reached where nearly all the documents are contained in only one cluster. It is therefore debatable

whether this type of clustering structure has any utility for retrieval purposes and calls the use of single-link and other related methods into question.

Croft (1977) has suggested that to overcome the inefficiencies of hierarchic clustering methods, in particular single link, core clustering may be used. This involves clustering a sample of the collection hierarchically, and then assigning the remaining documents to the resulting clusters on a heuristic basis. As he rightly points out, this is very much dependent on the initial choice of the core. This research indicates that the core should consist of a large number of documents in order to obtain anything like a reasonable clustering structure.

5.4 APPLICABILITY TO OTHER COLLECTIONS.

An important aspect of this research is its applicability to sets of documents other than MEDLARS. Two extremes are considered:-

INSPEC.

INSPEC documents are very similar to their MEDLARS counterparts, in that they have both been extracted

from a much larger operational database and have roughly the same average number of terms per document, and similar cluster statistics. It can therefore be expected that if the same experiments were performed using INSPEC data instead of MEDLARS, similar results would be obtained. Thus, the INSPEC test collection of only 541 documents may well not be representative of the full INSPEC collection, and any experimental results may be doubtful.

Cranfield 2.

The Cranfield 2 collection is so markedly different from MEDLARS and indeed most other collections, that it is difficult to associate it with them at all. The Cranfield 1400 collection is not made up of a subset of a larger collection, but was specially created as a collection in itself for use solely in IR experiments. Furthermore, its unusually high number of terms per document makes it a special case, for example in its clustering behaviour.

Indeed, because of its richness of terms and the restricted nature of its documents, Cranfield may even be able to provide a better prediction of retrieval behaviour in an operational environment than a large subset. Indeed, in the absence of firm conclusions

from the present research as to the size of collection required for meaningful results, the author would suggest that Cranfield provides the best alternative as a test collection, particularly when the cost of experimentation is taken into account.

5.5 CONCLUDING REMARKS.

This research has not been able to give any more than broad indications as to the exact size of collection subset necessary to ensure that retrieval performance is accurately predicted in an operational environment. Nevertheless, the conclusion of this work is that much larger collections than have previously been used in IR experiments are required if satisfactory results are to be guaranteed. This does not exclude the use of small collections as tools for the development and testing of IR systems but as soon as the point is reached at which an evaluation of the effectiveness of the system is required, they should be abandoned in favour of larger, more representative collections.

6 SUGGESTIONS FOR FURTHER RESEARCH.

The research reported in earlier chapters has uncovered areas in which further work may prove significant within the scope of information retrieval in general.

6.1

The most striking of these is concerened with document clustering. The results of the clustering experiments in this research indicate that the single-link method may not perform as well as it has been claimed and that previously favourable results may have been affected by the choice of test collection, namely Cranfield 2. The clustering of the MEDLARS data, although it has been carried out on a small collection of documents (only 5,000 but significantly more than some previous experiments (Jardine and van Rijsbergen (1971), van Rijsbergen and Croft (1975))) tends to be of an "all or nothing" nature. Either there are a large number of small clusters containing 2 or 3 documents at low dissimilarity levels, where the number of documents incorporated into the hierarchy is unacceptably low, or at a level where a more satisfactory percentage of documents are clustered, they tend to be members of a small number of very large

clusters. This continues until the point is reached where all the clusters become amalgamated into a single cluster and the only documents that are excluded are not clustered with any other documents at all.

The situation is unlikely to improve with larger collections. The work on combinations which examined collections of up to 25,000 documents, showed that combinations generally occur with a very low frequency and therefore the number of documents with terms in common is also low. This does not bode well for a clustering technique based on this figure.

The utility of this type of clustering in terms of retrieving documents is questionable. Either a large number of documents are retrieved or only two or three. There appears to be no position for compromise.

This is an area where further research is necessary in order determine the cause of such an effect and a suitable remedy. It would be desirable to be able to incorporate more documents into the cluster hierarchy at lower levels so that reasonably sized clusters could be formed. It is difficult to see how exactly to achieve this as single-link has the most relaxed requirement possible: for a document to be included in a cluster it needs only to

have one term in common with only one other document in that cluster. A strengthening of this criterion would certainly have the effect of preventing, or at least postponing the formation of the large cluster, but will inevitably lead to a dramatic reduction in the percentage of documents in the hierarchy.

6.2

In the light of the somewhat unique behaviour of the Cranfield 2 test collection which manifests itself particularly in the clustering experiments, the following questions arise:-

1. The Cranfield collection is very rich with regard to the number of index terms assigned to each document. What is the effect of enriching a collection by increasing the number of terms per document? Experiments which are able to compare Cranfield with for example, an enriched version of an MEDLARS/INSPEC-type collection could decide whether Cranfield is a better research tool than a representative subset from a large collection.

2. The Cranfield documents constitute a relatively

restricted subject area (aeronautics) and as such random subsets taken from so small a main collection (only 1400 documents) are unlikely to be useful in retrieval experiments. This prohibits experiments of the type performed in this research. Nevertheless, because of its narrow subject area it may be possible that it can be treated as a 'full' collection in its own right, subject to satisfactory experimental confirmation. With regard to operational collections such as MEDLARS and INSPEC, what is the effect of scaling down a large collection covering a wide subject area in order to produce a subset restricted solely to one small subject area and would this give a sensible prediction of the retrieval behaviour of the large collection? This can be investigated experimentally by extracting a section of the collection by performing a broad search covering a particular subject area and using the retrieved documents as the test collection.

6.3

It is unfortunate that this work has not led to more conclusive results but like so much research it is

justified by the ideas for further investigation that it has uncovered.

References.

Aitchison et al (1970)

Aitchison, T.M., Hall, A.M., Lavelle, K.H. and Tracy, J.M., "Comparative Evaluation of Index Languages", Part 1, Design, Part 2, Results, Project INSPEC, Institution of Electrical Engineers, London (1970)

Barber et al (1972)

Barber, A.S., Barraclough, E.D. and Gray, W.A., "MEDLARS on-line search formulation and indexing", Technical Report Series No. 34, Computing Laboratory, University of Newcastle upon Tyne (1972)

Barker et al (1972)

Barker, F.H., Veal, D.C. and Wyatt, B.K., "Comparative Efficiency of Searching Titles, Abstracts and Index Terms in a Free-Text Data Base", Journal of Documentation 28, 22-36 (1972)

Barraclough et al (1975)

Barraclough, E.D., Hunter, J.A., Lovett, A.J. and Rossiter, B.N., "the Medusa Current Awareness Experiment", Technical Report Series No. 78, Computing Laboratory, University of Newcastle upon Tyne (1975)

Burton and Kepler (1960)

Burton, R.E. and Kepler, F.W., "The Half-life of some Scientific and Technical Literatures", American Documentation 11, 18-22 (1960)

Cleverdon et al (1966)

Cleverdon, C.W., Mills, J. and Keen, M., "Factors determining the Performance of Indexing Systems", Vol. 1 Design, Vol. 2 Test Results, ASIIB Cranfield Project, Cranfield (1966)

Cleverdon (1972)

Cleverdon, C.W., "On the inverse relationship of recall and precision", Journal of Documentation 28, 195-201 (1972)

Cox and Dews (1967)

Cox, N.S.M. and Dews, J.D., "The Newcastle File Handling System",
in "Organisaton and Hardling of Bibliographic Records by Computer" (Edited by N.S.M Cox and M.W. Grose), Oriel Press (1967)

Croft (1977)

Croft, W.B., "Clustering large files of documents using the single link method",
Journal of the American Society for Information Science 28, 341-344 (1977)

Date (1981)

Date, C.J., "An introduction to database systems",
3rd. Edition, Addison-Wesley System Programming Series (1981)

Gibbs (1977)

Gibbs, M.E., "The examination and evaluation of the discrimination value method as applied to the MEDLARS document collection",
M.Sc. Dissertion, University of Newcastle upon Tyne (1977)

Hall (1977)

Hall, J.L., "Online Information Retrieval 1965-1976", ASIIE bibliography No. 8, ASLIE, London (1977)

Houston and Wall (1964)

Houston, N. and Wall, E., "The distribution of term usage in manipulative indexes", American Documentation 15, 105-114 (1964)

Jardine and Sibson (1971)

Jardine, N. and Sibson, J., "Mathematical Taxonomy", Wiley, London and New York (1971)

Jardine and van Rijsbergen (1971)

Jardine, N. and van Rijsbergen, C.J., "The use of hierarchic clustering in information retrieval", Information Storage and Retrieval 7, 217-240 (1971)

Keen and Digger (1972)

Keen, E.M. and Digger, J.A., "Report on an Information Science Index Languages Test", Aberystwyth College of Librarianship, Wales (1972)

Lancaster (1968)

Lancaster, F.W., "Information Retrieval Systems: Characteristics, Testing and Evaluation", Wiley, New York (1968)

Lynch (1977)

Lynch, M.F., "Variety Generation - A Reinterpretation of Shannon's Mathematical Theory of Communication, and its Implications for Information Science", Journal of the American Society for Information Science 28, 19-25 (1977)

National Library of Medicine

National Library of Medicine, "The Principles of MEDLARS", U.S. Department of Health Education and Welfare.

Olive et al (1973)

Olive, G., Terry, J.E. and Datta, S., "Studies to compare retrieval using titles with that using index terms", Journal of Documentation 29, 169-191 (1973)

Robertson (1975)

Robertson, S.E., "A Theoretical Model of the Retrieval Characteristics of Information Retrieval Systems", Ph.D. Thesis, University College, London (1975)

Robertson (1981)

Robertson, S.E., "The methodology of information retrieval experiments",
in "Information Retrieval Experiment" (Edited by K. Sparck Jones), Butterworths, London (1981)

Rocchio (1966)

Rocchio, J.J., "Document Retrieval Systems - Optimisation and Evaluation",
Ph.D. Thesis, Report ISR-10, Computation Department, Harvard University (1966)

Salton, Yang and Yu (1975)

Salton, G., Yang, C.S. and Yu, C.T., "A theory of term importance in automatic text analysis",
Journal of the American Society for Information Science 26, 33-44, (1975)

Sibson (1973)

Sibson, N., "SLINK: an optimally efficient algorithm for the single-link cluster method",
The Computer Journal 16, 30-34 (1973)

Sparck Jones (1973)

Sparck Jones, K., "Collection Properties Influencing Automatic Term Classification",
Information Storage and Retrieval 9, 499-513, (1973)

Sparck Jones and van Rijsbergen (1975)

Sparck Jones, K. and van Rijsbergen, C.J., "Report on the need for and the provision of an 'Ideal Test Collection'",
BLRDD Report 5266, Computer Laboratory, University of Cambridge, (1975)

Sparck Jones and van Rijsbergen (1976)

Sparck Jones, K. and van Rijsbergen, C.J., "Information Retrieval Test Collections",
Journal of Documentation 32, 59-75 (1976)

Sparck Jones and Bates (1977)

Sparck Jones, K. and Bates, R.G., "Report on automatic indexing 1974-1976",2 vols.,
BLRDD Report 5428, Computer Laboratory, University of Cambridge (1977)

Sparck Jones (1981)

Sparck Jones, K., "Retrieval system tests 1958-1978",
in "Information Retrieval Experiment" (Edited by K. Sparck Jones), Butterworths, London (1981)

van Rijsbergen (1971)

van Rijsbergen, C.J., "An algorithm for information structuring and retrieval",
The Computer Journal 14, 407-412 (1971)

van Rijsbergen and Sparck Jones (1973)

van Rijsbergen, C.J. and Sparck Jones, K., "A test for the separation of relevant and non-relevant documents in experimental document collections",
Journal of Documentation 29, 251-257 (1973)

van Rijsbergen and Croft (1975)

van Rijsbergen, C.J. and Croft, W.B., "Document clustering: an evaluation of some experiments with the Cranfield 1400 collection",
Information Processing and Management 11 , 171-182 (1975)

van Rijsbergen (1979)

van Rijsbergen, C.J., "Information Retrieval",
2nd. Edition, Butterworths, London (1979)

Vaswani and Cameron (1970)

Vaswani, P.K.T. and Cameron, J.B., "The National Physical Laboratory Experiments in Statistical Word Associations and their Use in Document Indexing and Retrieval",
National Physical Laboratory (1970)

Willett (1981)

Willett, P., "A fast procedure for the calculation of similarity coefficients in automatic classification",
Information Processing and Management 17, 53-60 (1981)

Williams (1980)

Williams, M.E., "Database and Cnline Statistics for 1979", ASIS Bulletin 7,(December 1980)

Winograd (1972)

Winograd, T., "Understanding Natural Language", Edinburgh University Press, Edinburgh (1972)

Appendix 1: List of Check Tags.

adolescence	adult
aged	animal experiments
case reports	cats
cattle	child
child, preschool	clinical research
comparative study	current biography-obituary
dogs	female
guinea pigs	historical article
historical biography	history of medicine
ancient	15th Century
16th Century	17th Century
18th Century	19th Century
20th Century	medieval
modern	human
in vitro	infant
infant, newborn	male
mice	middle age
pregnancy	rabbits
rats	review