1982

Probabilistic Models For The Simulation Of Bibliographic Retrieval Systems

Michael John Nelson

PROBABILISTIC MODELS FOR THE SIMULATION OF BIBLIOGRAPHIC RETRIEVAL SYSTEMS

by

Michael J. Nelson

School of Library and Information Science

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario

November, 1981

ABSTRACT

A general model of a bibliographic retrieval system is presented which has five main elements: the documents, the queries, the thesaurus of indexing terms, the search algorithms and the physical storage locations. This is adapted to produce a probabilistic model which is suitable for simulation purposes, concentrating on the assignment of index terms to documents. This is accomplished by using the distribution of terms over documents and over queries, the distribution of exhaustivity over documents and over queries, the distribution of co-occurrences (occurrences of pairs of terms), the distribution of relevant and non-relevant documents over the number of terms matching the query. Several theoretical distributions were tested against four databases to find the best fitting distributions using the chi-square criterion. The distribution of terms over documents was split into two parts. The low frequency terms were analyzed using the number of terms which occurred x times, called the frequency-size approach. The high frequency terms were ranked by the number of occurrences in documents and analyzed using the rank versus the frequency of the term, called the frequency-rank approach. It was found that a generalized Zipf distribution fit the frequency-size portion

and a generalized Bradford or log-rank distribution was best for the frequency-rank part.

These distributions were incorporated into a simulation program using a probabilistic model of term occurrences and co-occurrences. Simulation of the four databases was carried out using both the independence assumption of the occurrence of terms and the dependence assumption. In most cases the dependence model gave an improvement over the independent model but did not reproduce fully the original distribution of co-occurrences.

A small experiment with the clustering of terms to incorporate term dependence was also carried out. A method of incorporating the clustered terms into a simulation model needs to be found.

More work needs to be done in incorporating dependence of index terms, especially of order higher than two, into a model of bibliographic retrieval systems. Goodness-of-fit tests and parameter estimation methods need to be devised for the type of long tailed distributions encountered.

ACKNOWLEDGEMENTS

I would like to thank everyone at the School of Library and Information Science, The University of Western Ontario for their support in this project. Also those people who served on my advisory committee, especially Professor Jean Tague, my principle advisor who willingly gave of her time and energy to encourage and help with this research. A special thanks to my wife Claire and the rest of the family who supported my efforts throughout the research and writing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The use of computerized bibliographic retrieval systems in libraries and information centres in the last ten years has increased to a very high level, with thousands of users, hundreds of databases and many commercial vendors of systems to access the databases[126].

These bibliographic retrieval systems store document references in sets of integrated files (or databases). The user with an informational need (query) retrieves sets of references by the use of keys, either words from the title or abstract, index terms assigned by an indexer, or other attributes of the record, such as author, journal, year of publication.

There has been concern about the effectiveness of such systems and a lot of work has been done on experimental systems to evaluate the factors which affect performance, particularly the type and quality of indexing, and various characteristics of the query. Most of these experimental projects, such as Cranfield[24], Salton's SMART system[88],

and the Comparative Systems Laboratory at Case Western Reserve University [94] were done on small files of documents, up to a few thousand, whereas some of the operational systems have millions of references. Most of the evaluation was based on the elusive concept of 'relevance' of the document to the query. A few tests of operational systems have also been carried out based on basically the same evaluation criteria, as for example in Lancaster's test of Medlars[61].

The research done here is concerned with a different approach. The techniques of mathematical modelling and simulation will be used as a first step to evaluating systems.

Mathematical modelling is an iterative process which can be thought of as a four stage cycle (see Figure 1), following Roberts[85]. In the first stage real world data about bibliographic retrieval systems is gathered; in the second stage a mathematical description or summary of the findings is formulated, often with simplifying assumptions. The next two stages are concerned with testing the validity of the model. In the third stage mathematical predictions are made from the model and and translated into real world predictions. In the final stage the results are compared to the real world data which could be the original observations or new observations (i.e. testing the predictive power of

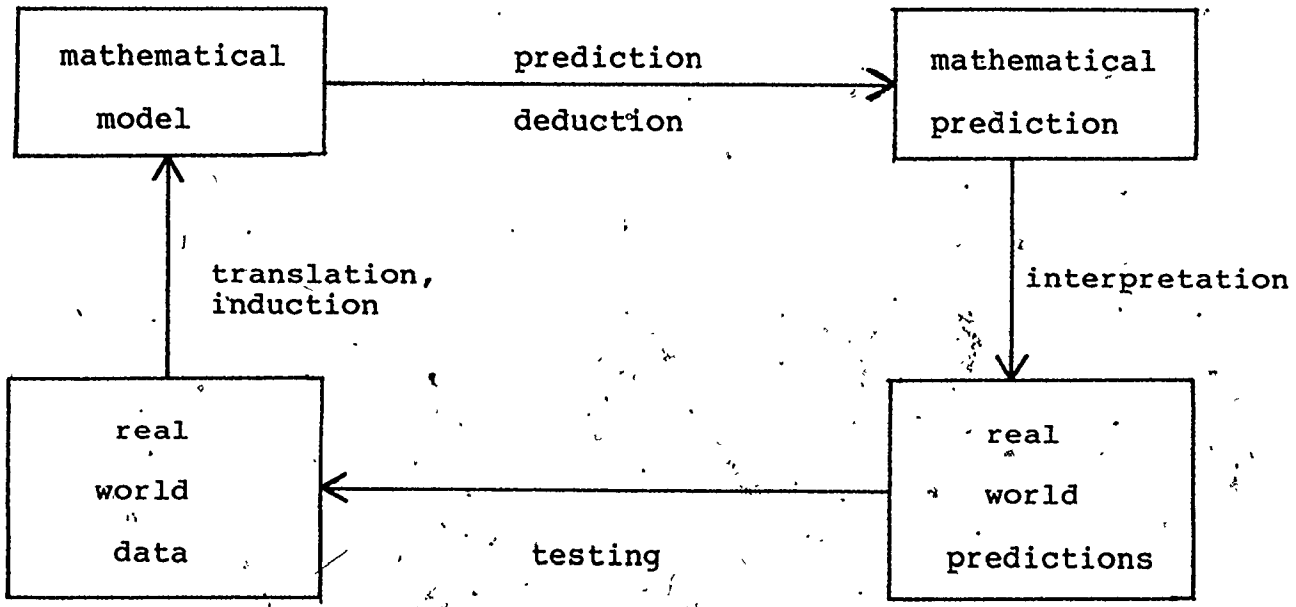

mathematical model
prediction
deduction
mathematical prediction
translation, induction
interpretation
real world data
testing
real world predictions

Figure 1.

the model). This completes the cycle, which can be repeated by noting discrepancies between the model and the real world data, then modifying the original model to make improvements.

According to Zeigler[131], there are three levels of validity. The first is replicative validity, where the data from the model matches the data already acquired and used to build the model. Secondly, the model is predictively valid if it can match data acquired after the model makes predictions. Finally, a model is structurally valid if it not only matches the real system data, but also produces the data with the same mechanism as the real system, so that we are modelling the internal structure of the system.

Another aspect which is important to this study is the relationship of simulation to mathematical modelling. Again Zeigler[131] says that simulation is essentially the computer implementation of the model, a way to produce predictions given a specific model. This is particularly useful, and many times essential, for probabilistic models, where the computer can sample probability distributions and perform the many repetitions and calculations needed to make predictions from the model. We can also easily change the basic parameters of the model to make predictions for various real world systems. In bibliographic retrieval systems such parameters might be vocabulary size, documents

set size, search procedure, and file structures. It is much easier to experiment with the parameters in the simulation model than in real systems.

Many models have been proposed, but very few have been tested for validity, even in the simplest cases. Each of these models works on the same basic document retrieval model of m documents and n keywords (or attributes) which can be represented by an m by n matrix X, where X(i,j)=1 if the ith document has been assigned the jth keyword; X(i,j)=0 otherwise. To represent the weighting of index terms, we can assign values of x(i,j) between 0 and 1.

When this system is modelled in a probabilistic manner, the simplest model is the standard multivariate one, with each of the n keywords being a binary variable, pairwise independent. The independence assumption has been made in many of the probabilistic models (e.g. Yu and Salton[130], Robertson and Sparck-Jones[86]) because it makes the mathematics more tractable. This assumption has not been tested rigorously and widely. To make the model more realistic it may be necessary to take into account pairwise dependence, then dependence in threes, etc. Some efforts in this direction are reviewed in section 3.2.4.

If we are interested in the distribution of document vectors X(I)=(X(I,j)) j=1,n for document I, the "curse of

dimensionality" as referred to by Van Rijsbergen (p115 [119]) becomes important. Then each term is considered as a separate dimension and a document is located in this n-dimensional space by the terms which are assigned to it. This means that there are, probabilistically speaking, a very large number of possible combinations of keywords which may occur (2^n), but in a particular document set we have only a relatively small sample of these combinations. There are no more than m actual combinations present, one for each document.

The dimensionality causes problems in calculating a goodness-of-fit test for a candidate distribution of index terms and their co-occurrences. The multivariate model and the dependence assumption also pose computational complexity questions because products and sums must be taken over many dimensions.

The purpose of this research is to gain greater understanding of bibliographic retrieval systems by the building of general probabilistic models which overcome some of these problems, to develop a methodology for testing the models against sample data, and then to use the models in simulating document retrieval systems of various types. The simulation will use the mathematical model to describe, by the use of a small number of parameters, the general characteristics of the keyword index, and then to describe

or predict some aspects of the original retrieval system. In particular, the independence model of keyword assignment will be tested, and an attempt made to determine the degree of dependency required. Methods of incorporating the dependency information into the model, if needed, will be investigated.

The specific problems approached are the simulation of the document-term matrix and the simulation of the user's judgement of the relevance of the documents to a query. The accuracy of the model in describing the document-term matrix is evaluated by comparing the frequency distribution of term occurrences and the frequency distribution of term co-occurrences from the model with the original data and data from the independent model. The term occurrence models are an explicit part of the basic model but the frequency distribution of co-occurrences are not directly incorporated into the model so form a good test of validity. Predictive validity is checked by predicting the parameters of the distributions of term occurrences for the second half of a large database from the occurrences in the first half.

The user judgement model is evaluated first by comparing the frequency distribution in the model of the number of query terms matching in relevant documents with the corresponding empirical distribution by means of the chi-square test. This was also done for the distribution of

the number of query terms matching in non-relevant documents. Then a comparison of the graphs of the recall-precision curves for the real and simulated data is done and a two-sample Kolmogorov-Smirnov test to compare the recall versus cosine similarity distribution between the simulated and real sets is carried out. If the recall-precision curves match and the recall versus cosine similarity curves match in the simulated and real data, this is strong evidence that the model is reproducing the original system. In each case parameters for the distributions used in the model are estimated from the original data. When testing the user judgement model, the original document-term matrix is also being tested as it is part of the complete model.

CHAPTER 2

A GENERAL MODEL

To provide a framework for the discussion of other research and the specific models used in this work, the following general mathematical model will help to discriminate the various components of a bibliographic retrieval system. This model was first presented by Tague and Nelson[111].

A bibliographic retrieval system can be modelled by an 11-tuple,

$$B=<D,Q,T,A,P,i,r,s,a,b,c>$$

where the capitals represent sets and the lower case letters represent functions. The sets are:

D is a set of document descriptions, where a document is any form of recorded information-- book, journal article, report, film, audio tape, etc.

Q is a set of user query descriptions, where a query is any verbalized information need.

T is the set of terms or keywords, possibly having a structure as in a thesaurus.

A is the set of search algorithms.

P is the set of physical storage locations, which might be in a computer.

Each of these sets may have various mathematical structures defined on their members to represent a particular kind of organization of these members. Some of the structures used here are S(X), the power set of X (i.e. the set of all subsets of X), the set H(X) of all boolean expressions formed from the members of X, the set O(X) of all rankings or preference orderings of the elements of X, the nth product set X^n of X (i.e. the set of all n-tuples whose elements are members of X), the set V(X) of all tuples from X (i.e. $V(X)=\bigcup_i X^i$), the set G(X) of all trees formed from the members of X, and $C_n(X)$ the set of all n-way partitions of X. The symbol M(X) will be used to represent a set containing some variety of discrete mathematical structures defined on X.

The functions describe relationships among the various sets in the model, and in general describe a process which is carried out in the system. Generally, these functions can be described as follows:

i: D -> M(T) is the indexing function, which assigns a set of terms from the term set to each document.

r: Q -> M(D) is the retrieval function, which assigns a document set structure (e.g. ordering, partitioning into relevant and nonrelevant) to each query.

The retrieval function may be further analyzed into the composition of four subfunctions:

$$r = r_1 \cdot r_2 \, r_3 \, r_4$$

r_1 : Q -> M(T) is the search strategy construction function, which assigns some structure of terms to each query.

r_2 : M(T) -> A is the search algorithm selection function, which selects an encoded procedure for searching the physical locations in which the document representations are stored.

r_3 : A -> V(P) is the physical access function, which determines the sequence of physical storage locations which are accessed by the search algorithm.

r_4 : V(P) -> M(D) is the output function, which outputs some structured set of document representations, on the basis of the information found at the various physical locations accessed.

s:T$\cup$D -> S(P) is the storage function which assigns sets of physical storage locations to the elements of the document and term sets.

a: Q -> M(D) is the assessment function, which assesses the document set with respect to the query by imposing some structure on it (e.g. partitioning into relevant and nonrelevant sets, ordering by relevance, assignment of a relevance weight).

b: O_1 x O_2 -> F(R) is the search evaluation function where $O_1=\{u:u=r(q),\ q\in Q\}$ and $O_2=\{u:u=a(q),\ q\in Q\}$ are the retrieval outputs and F(R) is the set of all real valued functions.

c: AxP -> R is the complexity function which associates a real number with an algorithm and physical storage assignment. This complexity may be further analyzed as internal and external complexity, relating respectively to the algorithm and the storage data structure.

CHAPTER 3

REVIEW OF PREVIOUS RESEARCH

3.1 STATISTICAL INDEXING

Statistical automatic indexing is mentioned here because it implicitly uses a model of information retrieval based on the frequency characteristics of words in text for determining keywords or class numbers.

The advent of digital computers in the 1950's stimulated an interest in automatically indexing documents for subject retrieval by using full texts or abstracts in machine readable form. There are two basic methods of indexing documents automatically: assignment indexing and derivative indexing. Assignment indexing attempts to assign either a classification code or index term from an existing thesaurus by analyzing the text. In derivative indexing the index terms are chosen from the words used in the document.

Two of the earliest attempts at derivative indexing were by Luhn[67] in 1957 and Baxendale[5] in 1958. The basic technique was to count the number of times words

occurred in the text. The most frequent were function words such as "of", "by", "in", with no content. The most frequent non-function words in the text were taken as representing the subject content of the document. Luhn also used pairing of words in sentences and the positions of the words in the sentences as clues to indicate importance. Baxendale compared the method of using basic frequency information to choose index words with two other methods which used syntactic information. The first restricted frequency counts to words which occurred in the first and last sentence of a paragraph, assumed to be the topic sentences. The second method counted only words which occurred within four words of a preposition, hopefully within a prepositional phrase. The evaluation of these methods proved inconclusive.

Using co-occurrence and association measures for selecting index terms implicitly says that the terms are not independent.

There have been many studies in the area of automatic indexing, many of which are reviewed in Liebsney[64]. The use of syntactics and semantics has continued in automatic indexing research, but the main focus of this study is the statistical method, and linguistic techniques will be mentioned only when they interact with these methods.

3.2 MODELS IN INFORMATION RETRIEVAL

The first probabilistic model to incorporate these ideas in the complete information retrieval situation was the 1960 paper by Maron and Kuhns[72]. Their study was based on the estimation of a relevance number for each document. The relevance number was defined as the probability that if a user requests information on a specific term he will be satisfied with the document. The goal was to produce an ordered list of documents for each request, ranked according to relevance. This is essentially a weighting scheme, where the terms are weighted by the relevance number and the relevance number is estimated from weights given to the index terms by the indexers. This scheme depends heavily on the indexer estimating relevance without having a precise definition of this elusive concept and without a specific query.

Maron and Kuhns also propose augmenting the query with terms chosen by means of a closeness measure between two terms. This is calculated from the frequencies of occurrence and co-occurrence of the two index terms and is essentially the degree they vary from statistical independence, so Maron and Kuhns are assuming that the terms are dependent. They also propose using document distance or closeness to add to the retrieved document set. These closeness measures are the basis for clustering and

automatic indexing of documents using statistics of words from the documents or their abstracts.

The idea of using the association of terms was next investigated in 1961 by Stiles[104] and carried on by many others in the next few years. Many of these results were reported in the Symposium on Statistical Association Measures for Mechanized Documentation[103]. The majority of papers in the symposium were concerned with measures of association between words or index terms. A few studied the properties of the association measures and others applied these measures to term selection or classification. Another feature was the several papers proposing solutions to the "dimension problem" mentioned previously. Various statistical matrix techniques such as factor analysis (Borko[13]), discriminant analysis (Williams[125]), and latent class analysis (Baker[3]) were tested on small document sets. These ideas for dimension reduction have not been widely used, mainly because they are computationally inefficient and not practical for large data sets.

3.2.1 Descriptive Models

Swanson [106] developed a descriptive model of one type of retrieval system where the system ranks documents using the number of index terms the query has in common with the

document. This is called a z-cutoff search strategy and gives a quasi ordering to the set of documents. Swanson partitioned the document set for each query Q first by the number of terms assigned to the document, then by the number of these terms which matched with the query Q, and finally by relevance. The partition P is then:

$P = \{ N(p,j,i) \}$

p index terms per document

j terms in common with Q

i=0 irrelevant i=1 relevant

This is a partition since if a document has p index terms, j matching with Q, and we assume it is either relevant or not, the document must be in exactly one set N(p,j,i).

Swanson used this model to show the effects on precision and recall of randomly deleting index terms from documents. Although there have been more deletion experiments with test collections (e.g. Svenonius[105]), the model has not been used since the original paper.

Another more elaborate descriptive model was developed by Soergel[100], based partly on the earlier work by Mooers[72], Fairthorne[36] and Hillman[52] on mathematical theories of classification. He shows the attributes form a free distributive lattice and shows how roles, links, weighting of terms, associative retrieval and other features of retrieval systems can be described with the model. He

states in summarizing that this structural model should be a sound basis for the solution of functional problems in information storage and retrieval. There has not been much interest in this model since that time.

Salton summarized the Hillman and Fairthorne models in his 1968 book [90], where he gives the elements of set theory and lattice theory and describes some of the models used.

Turski's model [116] evolves in the same direction by defining a retrieval system by a quadruple (T,D,R,r) where T is the thesaurus, D a collection of documents, R a query set, and r is a mapping $r:R->2^D$ which maps a query to a subset of the documents. Turski develops the thesaurus properties of T by defining an equivalence relation (synonymy) on the set of documents and a generalization relation R. He uses the thesaurus relation R to extend the usual definition of inclusion of queries and extends the 'inclusive' retrieval system (see Salton[90]) to a thesaurus based system.

Marek and Pawlak [70] have also reported an algebraic model for information retrieval. The main structure used is a Boolean algebra for the index terms with other structures added. The main application is the development of certain properties of retrieval when documents are stored in

"generalized components" or clusters. Laus and Dabrowski[30] took the Marek and Pawlak model and added a hierarchical classification to the set of index terms. Mazur[73,74] extended this thesaurus based model by combining it with some of the features of Turski(see above). He adds weights to each term in the document descriptions and query descriptions.

A more recent, but simpler structural model is that of Bookstein and Cooper[10]. They give a very general set theoretic structure to an ISR system and show, with some examples, how particular features can be described. This seems to be only a first step in finding a model which can be used to gain further insights to the problems of information retrieval.

3.2.2 Fuzzy Set Theory Models

Two other papers also use the basic quadruple (T,D,R,r) but add a fuzzy relation to the model. Radecki's[80,81] fuzzy relation is based on a function $\mu_R(x,t):(D \cup R) \times T \rightarrow [0,1]$ which for each pair (x,t) gives a value between 0 and 1, where x can be in either the query or the document set and t is in the thesaurus. He also uses a function $\mu_S(t,t'):T \times T \rightarrow [0,1]$ and a generalization relation to develop the concept of a fuzzy thesaurus. Note that μ_R can be

interpreted as giving a weight to each term for each document and each request, so it is a generalized weighting function. μ_S can be interpreted as the degree of association between two terms. He goes on to show μ_R induces a quasi ordering on the requests and that the system satisfies the inclusion property (see Turski[116] and section 3.2.1).

Tahani[113] also uses fuzzy relations by defining a function g:R x D -> [0,1], a "matching index" between a document and a query, and a function h:D x T -> [0,1], the index weight of a term in a document. These functions are used to define fuzzy sets in the usual way. An algorithm for retrieval using Boolean queries which retrieves a fuzzy set of documents is described.

3.2.3 Vector Space Models

Another series of models is based on the concept of a vector space. The basic idea is to use each of the n keywords as a coordinate in an n-dimensional vector space, where each document is given the coordinates in the space corresponding to the weights of each of the terms, which in many cases is binary. We can use the properties of vector spaces to define Euclidean distance between documents, etc.

Takahama [114] uses such a model and adds query vectors and definitions of "vagueness" for a query and "similarity" between queries and documents. Frequency of occurrence and co-occurrences of terms is used in calculating the similarity between document vectors and Boolean queries. Matrix methods are used throughout for representations and calculations. Algorithms for implementation in an SDI system are given but no retrieval results are included.

Salton and his students have also used a vector space model as the basis of many investigations using the SMART system. He has studied many aspects of information retrieval systems including correlation coefficients, clustering of documents and terms, thesaurus construction, automatic indexing, feedback and term discrimination. The results of the indexing experiments are summarized in Theory of Indexing and Dynamic Information and Library Processing. The basic model is a vector space of n dimensions, where n is the number of index terms, with the cosine coefficient as a measure of closeness between document vectors. The term discrimination model uses the density of the document space as defined by the average document similarity as the basis for a measure of the effectiveness of the index term. If the removal of the term makes the average document similarity decrease then the term is a better "discriminator". The assumption being made is that "... a

document space which is 'bunched up' in the sense that all documents exhibit somewhat similar index vectors is not useful for retrieval since one document cannot then be distinguished from another..."[89,p8].

The discrimination value is used to weight terms in an information retrieval experiment and is compared with other weighting schemes using the usual recall/precision measurements. This experiment indicates that, on the average, discrimination value weighting gives better retrieval than other types of frequency based weighting. It must be remembered that we cannot evaluate individual terms on this basis, as we are only using averages.

Another vector space model is that of Cleveland[23] who combines Goffman's indirect retrieval[41] with a vector space in order to represent distances between documents in terms of Euclidean distances. He uses a four dimensional space with four measures of document distances as the four dimensions. The model is used for calculating a combined document distance measure.

3.2.4 Probabilistic Models

One of the earliest probabilistic models following Maron and Kuhns was by Miller[75]. He gives each search term a weight $W = \log(w/p)$ where p is the proportion of the

documents in the system indexed by the term, and w is the user's estimate of the proportion of relevant references to which the term would apply. These proportions can be interpreted as probabilities, p being the probability a document is indexed by the search term, w the probability a relevant document is indexed by the search term. In terms of Shannon's information theory W is the quantity of information about relevance transmitted by the occurrence of the search term. In developing the weight of a coordinate set of search terms, Miller assumes the independence of index terms so that the occurrence of one term cannot affect the probability of another. To use this in retrieval, the documents are ranked by the sum of the weights of the search terms which occur in their indexing. This technique was compared with a Boolean search on the Medlars data base and found to be an improvement, especially when the Boolean search retrieved a small number of references.

Yu and Salton[130] present a probabilistic model for evaluating automatic indexing which uses relevance criteria to weight the terms occurring in user queries as a function of the balance between relevant and non-relevant documents in which these terms occur. One of the basic assumptions used is that the assignment of index terms is independent.

The model by Robertson and Sparck-Jones [86] is the basic binary multinomial with the assumption of independence

of terms. Their main concern was to incorporate relevance information into the retrieval decision. They did, however, modify the total independence assumption by assuming independence of terms within relevant documents and within non-relevant documents. Croft and Harper [28] have recently modified the relevance feedback part of the model by assuming the top ranked documents in the initial search are the most relevant.

Van Rijsbergen [119] extended the relevance model by adding the dependence of each term on exactly one other term, thereby forming a tree structure of dependence using the Chow expansion [21]. This was tested by Harper and Van Rijsbergen [47] using a variety of relevance weighting techniques and dependence information. Further experiments with this model are reported by Van Rijsbergen, Harper and Porter[120].

Another model which incorporates term dependence is the linked-2-Poisson model by Bookstein and Kraft[10a]. The basic model is the 2-Poisson model of word occurrence in documents (see 3.3.1 and [11]). This is extended by incorporating term dependencies by means of the joint probability of s occurrences of one term and t occurrences of another term in a document. This is used to derive a decision rule for the retrieval of documents. In general there will be a large number of joint probabilities to

estimate from the co-occurrence frequencies.

Yu, Luk and Siu [128] have developed a model with an arbitrary number of term dependencies represented by means of the Bahadur-Lazarfield expansion. This is applied to a clustered search process and tested with the dependencies among the query terms in a clustered document space in the SMART system. In the same paper the computational complexity of the algorithms is also investigated.

Tague [108] and Nelson [78] have also investigated a form of the Bahadur-Lazarfield expansion for pairwise dependencies, but discovered problems in fitting data to the model because multivariate goodness of fit tests for sparse data were needed.

3.2.5 Summary

There have been three reviews of mathematical models in information retrieval in the last few years, by Robertson[87], Salton[91], and Yu, Luk, and Siu[128], all of which have tried to summarize and evaluate these many approaches to modelling retrieval systems.

In many of the above models, very little testing against data from retrieval systems was presented. Some of the models were tested by using the predictions from the

model to improve retrieval performance as measured by precision and recall. It was not the internal accuracy of the model, but the effect on the output which was measured, particularly for the probabilistic models.

3.3 DISTRIBUTION OF INDEX TERMS

Most of the models in the previous section do not use analytical descriptions of the distributions of index terms, but rely on taking empirical values for each term. If we can describe the index terms with a small number of parameters as in a distribution, we can more easily characterize a particular system. Distributions are also very useful in simulation for the same reasons. There are two types of distributions which are commonly used to summarize frequency data, the frequency-size and the frequency-rank as differentiated by Brookes and Griffiths [16] and Hubert [57]. The frequency-size distribution $g(x)$ is the number of terms which have x postings, $x=1,2,3...$, and the frequency-rank $f(r)$ is the number of postings of the r-th ranked term, $r=1,2,3...$, where the terms are ranked in descending order by the number of postings. In the following, N is the number of index terms, K is the total number of postings, $G(x)$ is the cumulative distribution of $g(x)$, and $F(r)$ is the cumulative distribution of $f(r)$.

3.3.1 Distributions Of Words In Text

There have been many attempts to describe the distribution of the occurrence of words in English language text. One of the most quoted is that of Zipf[133], who showed that $f(r)=c/r$ (c=constant) for a large number of different samples of text. Many writers have given many versions and explanations of Zipf's Law. Mandelbrot[68] generalized this to get a better fit and used $f(r)=c/(r+d)^b$ (b,c,d parameters). Good[42] had a frequency size model for x=1 to 15 where $g(x)=c/(x^2+x)$. Simon[98] developed a stochastic model to explain word occurrences and derived $g(x)=c/x^2$ (c= parameter). Many more versions and explanations for the distribution of word occurrences in text have been developed by those mentioned above and others.

Many studies of the distributions of index terms have naturally been influenced by Zipf's distribution and several researchers have assumed this type of distribution of index terms without any further investigation. For example, Avramescu[2] uses this as the starting point to derive a formula for "query size", or number of documents matching the query, as a function of indexing exhaustivity and total number of query terms. Raver[84] claims a Zipf distribution of index terms but then assumes a different exponential distribution of the form $f(r)=f(0)(f(0)/f(N))^{(-r/N)}$.

Toma[113] graphically fits Raver's distribution against the Euratom Thesaurus, recognizing that this is different from Zipf's distribution. Sparck Jones[101] claims the Zipf distribution of index terms to derive a weighting function proportional to log(f(r)). Schuegraph[95] also uses a Zipf type distribution of index terms to derive methods for compression of inverted files.

Bookstein and Swanson[11] take a slightly different point of view by trying to find the difference in the distributions of the number of word occurrences in abstracts of relevant documents and non relevant documents. They postulated a 2-Poisson model in which the distribution of a word was Poisson in each set of documents, but with different parameters. This was tested by Harter[48,49] who used the model for automatically choosing index terms.

3.3.2 Testing Distributions Of Terms

One of the earliest studies which collected statistics for the distribution of index terms was by Schultz, Schwartz and Steinberg[96] in 1961. They looked at graphs of the distribution of indexing exhaustivity (number of terms used to index a document), grouped g(x), F(r)/F(n) on semi-log paper, and a grouped frequency size distribution for pairs of terms, all for two different databases. They did not try

to fit mathematical distributions to these graphs, only describing their general shape.

Houston and Wall[55] fit a log-normal distribution for g(x), where log(x) is distributed normally, to nine indexes but found this distribution did not fit for the 5% most heavily posted terms. They used a strictly empirical approach and found a complicated set of parameters to describe the log-normal distribution. They also looked at the relationship between vocabulary size and total postings.

McBirney[75] does a brief comparison of the use of thesaurus terms in the Bureau of Reclamation SDI system to 1970 and the Defense Documentation Centre statistics to 1966. He draws a graph of the cumulative frequency of term usage G(x) versus the percentage of the total vocabulary, but does no curve fitting. His main observation is that 50% of the terms were used 13 or fewer times in five years and for the DDC data 50% were used 7 times or less in 13 years.

Bennett[6] fits a generalized Zipf distribution $f(r)=a/r^k$ to Medlars subject headings and INSPEC terms, but the fit is not good for large rank, infrequently occurring terms.

Doszkocs, Schultheisz, and Vasta[33] plot grouped x versus g(x) on a log-log scale and claim the distribution generally follows a Bradford-Zipf-Mandelbrot type

distribution as described by Fairthorne[37] without actually fitting any particular distribution.

Griffiths[44,45] looks at the distribution of index terms for incorporation into a stochastic model for use in simulation. She looks at the distribution of terms from several systems and claims they follow a generalized Bradford distribution developed by Brookes[16]: $F(r)=k\log[(a+r)/a]$ where $k=1/\log[(a+N)/a]$. She does not use goodness of fit tests such as chi-square or compare the results to other distributions.

Tague, Nelson and Wu[112] have investigated the distribution of index terms in the SMART system. They found that a mixture of a Zipf distribution of g(x) for low frequency terms and a log-rank distribution for high frequency terms gave a better fit than a number of single distributions for all terms. The chi square test showed the split distribution was a better fit than the Zipf for f(r), Zipf for g(x), Griffiths form of Bradford's law for F(r) and the negative-binomial distribution for f(r).

3.4 MODELS OF THE DISTRIBUTION OF INDEX TERMS

In contrast to the empirical approach in the previous section, there have been some theoretical explanations of term distributions attempted.

Zunde and Slamecka[135] use information theory to show that the optimum distribution of terms is $p(t)=(1-1/p)/(p-1)$, where $p(t)$= probability that a term will have t postings, p=average postings. There are several problems with this approach. First the terms are all considered equal in their information content and second, one of the assumptions is that the number of postings in the system remains constant. Brookes[15] has also commented on the derivation, but has not resolved the problem.

Hodes[53] also uses information theory to evaluate the usefulness of a term in retrieval. He defines discrimination of a term by the reduction in uncertainty. Then if we assume the independence of terms, maximum discrimination happens when each term is posted to half of the documents. Hodes uses this discrimination measure to evaluate new terms coming in to an existing system and does not try to make the whole system conform to the theoretical ideal system.

Another possibility is to develop stochastic arguments for the occurrence of index terms such as Bird[7] did for the distribution of indexing exhaustivity. Some suggestions in this direction have been made by Griffiths[43] who proposed that the generalized Bradford is generated by a very mixed Poisson process. Some of the work done in the area of word distributions in text may also be applicable

(e.g. Simon[98], Sichel[96] and Hill[50,51]).

3.5 DISTRIBUTION OF INDEXING EXHAUSTIVITY

The main study of the distribution of indexing exhaustivity was done by Bird[7], where it is shown that a simple Poisson distribution of indexing exhaustivity provides a fairly good fit for most of the information systems tested, but that a negative binomial gave a better fit in some cases. The log-normal distribution did not fit. Bird also develops a stochastic argument which shows that a Poisson distribution can be expected for a homogeneous system, and if several of these document sets are added together we can expect a mixed Poisson process. One such mixed process is the negative binomial, which is the reason it was tested against the systems by Bird.

Griffiths[43] shows that the the generalized Bradford which she uses for the distribution of index terms is also a mixed Poisson and claims that the distribution of indexing exhaustivity also follows such a distribution, but there are no goodness of fit tests reported, and in her simulation model she assumes a simple Poisson distribution of indexing exhaustivity.

3.6 DISTRIBUTION OF CO-OCCURRENCES

As mentioned in the section on mathematical models, index terms are often assumed to occur independently. What evidence is there against this hypothesis? Can we describe the distribution of occurrences of pairs in some way?

Again the paper by Shultz, Schwartz and Steinberg[95] seems to be one of the first to publish results in this area. They plot a grouped frequency-size graph of the number of pairs versus the number of uses for two databases. They observe that the shape of the curve is the same as for single terms.

The studies mentioned in section 3.2 which used the association of terms in databases were of course using co-occurrence information. Most of these studies did not look at the distribution of co-occurences, except for Switzer[107], who noted that the distribution of the number of co-occurrences of terms is hypergeometric if the terms are assumed to occur independently. He then used the probability of a particular pair occurring as the significance of the term association. This says that the more the actual number of co-occurrences differs from the theoretical co-occurrences under independence, the more significant is the association. He uses this information to get a subset of index terms to use as a basis for a vector

space model.

Many studies of retrieval systems have used term-term correlations based on co-occurrences, e.g. Lesk[63], Cagan[17], and Stiles[104], mainly for clustering of terms, as clues to the semantic relationships between terms.

Curtice and Jones [29] investigated co-occurrences of text words within an abstract by using f= the frequency of occurrences of a term and N= the number of different terms with which this term co-occurs. They then calculated R=f/N and plotted R against N for all terms. They found that R was a decreasing linear function of N. Those terms which were furthest below the regression line (low R for a particular N) were judged as good index terms by a team of indexers.

Jacquesson and Schieber[58] analyzed term-co-occurrences for the ISIS system of the International Labour Office. This system had 40,000 documents indexed with 1400 keywords. The correlation between two terms is measured by the ratio of the number of co-occurrences to the number of documents which contain one term or the other. The total possible number of correlations is 1400 x 1400 = 1,960,000 but only 221,886 were non-zero. They grouped the coefficients into a table of values and hypothesized the distribution is Poisson. They also looked at tables which

showed all the terms which co-occurred with one particular term.

Griffiths[45] investigated the frequency of co-occurrence of all terms with one particular term and found a similar distribution to the distribution of terms (i.e. generalized Bradford for the cumulative co-occurrence frequency). This was done for several terms as we potentialy have a different distribution for each term.

Tague, Nelson, and Wu[112] found some evidence that the distribution of the number of co-occurrences with a term, given its occurrence frequency, has the form of a Zipf frequency-size distribution.

3.7 SIMULATION STUDIES

Some early simulation studies such as Baker and Nance[4] and Chapman[20] considered the simulation of information retireval systems from the overall system point of view including such things as costs, personnel, response time, number of users and hardware. This study is not interested in the management of retrieval systems but in the internal retrieval characteristics.

Fried et al[39] developed a simulation model of indexing for use in the study of automatic document

classification. To get parameters for the simulation they analyzed a sample of a thousand documents with 1096 distinct descriptors. They state the distributions of postings per term is exponentially decreasing and the distribution of co-occurrences of terms with one particular term is also exponentially decreasing with no further explanation. A simulation program assigned terms to documents until the maximum number of postings for a term was exceeded. Then an adjustment of the indexing was done by shifting some of the postings to an unused term. This was repeated until all the terms had the correct number of postings. They also simulated generic posting of terms. One of the main problems, which was recognized by the investigators, is that the simulation process is very different from the actual process of choosing index terms.

Another simulation model by Cooper[27] had five components: a thesaurus generator, document generator, query generator, search routines, and evaluation routines. This model is very general with a number of parameters which can be varied, including the form of the term distribution, measure of similarity between terms and exhaustivity of indexing. The thesaurus component includes co-occurrence information by having a term-term association matrix. The actual simulations were run on a simulated thesaurus of 200 terms and 150 simulated documents with 22 different query

files. The input values for the frequency distribution of terms and other parameters were not based on actual data from existing systems. The output was evaluated by the number of documents retrieved.

Another simulation model developed by Griffiths[44,45] has been mentioned previously. The input parameters for the term distributions were calculated from six large databases, the query characteristics from two of these systems, and the relevance assessments from two other studies. She found the distribution of index terms in documents and queries was the generalized Bradford in both cases. The distributions of the number of retrieved documents, relevant retrieved documents, and non-relevant retrieved documents were lognormal however. Griffiths compared these findings with the Cranfield collections and the Information Sciences Index Languages Test and found the distributions were not the same as the operational systems.

These data were used to build a simulation model which first created a vocabulary by using the generalized Bradford distribution, then created document distributions by assuming a Poisson distribution of indexing exhaustivity, and randomly assigning terms until the correct number of terms was attained. The co-occurrence information was not included in the final model. Queries were generated in a similar manner and matched against the document

descriptions. Then the number of relevant documents was simulated by the lognormal distribution.

A similar approach is taken in this research, except the distributions found are different and are tested by goodness-of-fit tests, and the co-occurrence information is incorporated into the model.

CHAPTER 4

DISTRIBUTIONS USED IN THE SIMULATION MODEL

4.1 DISTRIBUTIONS IN THE GENERAL MODEL

One way of explicating the individual functions from our general model in Chapter 2 is to consider each function as a stochastic process and to implement this process by a probability distribution of the objects under consideration. The distributions used are the distribution of terms over documents and over queries, the distribution of indexing exhaustivity (the number of postings per document), and the distribution of relevance over documents, for a given query.

Then the indexing function i: D -> S(T) can be simulated by sampling first from the distribution of indexing exhaustivity to get the number of terms assigned to the document, and then from the distribution of terms repeatedly to get the set of terms for a particular document.

Similarly the query formulation function r_i :Q -> S(T) can be simulated, with different parameters for the

distributions of indexing exhaustivity and term assignments.

From a distribution of relevance for a given query we can simulate the assessment function a: Q -> M(D) by assigning the values relevant or nonrelevant to each document.

4.2 GOODNESS-OF-FIT

The aim of our research is to find distributions which best describe the functions of interest. The classical statistical procedure recommends that we start from a completely specified theoretical distribution for a population and test the null hypothesis that the empirical distribution, which is from a random sample of the population, is the same as the theoretical one. The two most commonly used tests for this goodness-of-fit hypothesis are the Chi square and the Kolmogrov-Smirnov (see Conover[26] or Horn[54]).

In the case of modelling bibliographic retrieval systems we are taking a particular example and trying to approximate and parameterize it using a theoretical distribution. This is not the same as taking a random sample from a population. We must also remember that the distributions are discrete frequency functions, being counts of occurrences of terms, documents, co-occurrences , etc.

4.2.1 The Chi-square Test

The advantages of the chi-square test are:

i. that it is based on frequencies

ii. that it can easily accommodate the case where the parameters are estimated from the data by reducing the degrees of freedom by the number of parameters estimated

iii. it is easily calculated

The disadvantages are:

i. that if the expected cell frequencies are too small the results may be invalid, and we have frequency distributions with long tails of small expected frequencies

ii. the value of the calculated statistic is affected by sample size, so that, given a large enough sample, the empirical data will always depart significantly from the theoretical distribution.

The first disadvantage can be partially overcome by grouping the cells, but the grouping tends to be arbitrary and does not give a true picture of the tail of the distribution. The second discrepancy occurs because the theoretical model is only an approximation to the population distribution.

4.2.2 The Kolmogrov-Smirnov Test

The Kolmogrov-Smirnov test was designed for ordinal level variables representing underlying continuous variables. The statistic calculated is the maximum distance between the observed and expected cumulative probability distributions. The two main disadvantages are:

i. it is designed for continuous variables

ii. it is not easy to modify the test for the case where parameters are estimated from the data.

4.2.3 Conclusion

Although some specialized goodness-of-fit tests have been devised (see Conover[26]), they are not applicable to this study. Therefore the chi-square statistic will be used in most cases as a relative measure of the goodness-of-fit of various alternatives for the theoretical distributions.

4.3 SOURCES OF DATA

Data for building and testing the model was obtained from four databases. The basic statistics of these databases can be found in Table 1. SPI is an index to

TABLE 1

DATABASE CHARACTERISTICS

	Medlars	Cranfield	SPI	CIJE	CIJE-2
Number of Documents (m)	450	424	503	12167	12168
Number of Terms (n)	4726	2651	1150	4157	4054
Total Postings	18304	22708	3102	101212	97922
Average Number of Postings per Document	40.68	53.56	6.2	8.32	8.05
H1, Highest Frequency of a Term in Frequency-size	48	111	25	199	211
H2, Number Terms in Frequency-rank	24	18	6	43	43
Average Number of Terms per Query	9.1	7.3			
Average Number of Relevant Documents per Query (g)	13.4	8.4			
Number of Queries	24	24	0	0	0

periodical articles developed by Tague[110] at the Unversity of Western Ontario. The MEDLARS and Cranfield databases are from the SMART system at Cornell University. CIJE is Current Index to Journals in Education, record numbers EJ215324 to EJ227490 from 1980, obtained from the ERIC Processing Facility on magnetic tape. To test the predictive capability of the basic distributions used, a second part of the ERIC database called CIJE-2, consisting of record numbers EJ227491 to EJ239658, was also used.

4.4 DISTRIBUTION OF INDEX TERMS

An index term is taken to mean any content word or combination of words which is treated as a single entity in the indexing of documents. In the SPI database it was any word in the title or supplementary keyword field provided by the users. For CIJE a term was a descriptor from the descriptor field (field 35) of the ERIC tape record, which can be more than one word, and is assigned by an indexer from the ERIC Thesaurus. The Cranfield and MEDLARS terms are also the keywords or descriptors assigned to a document by indexers.

As explained in 3.3, there are two basic approaches to the distribution problem, the frequency-rank approach

and the frequency-size. If we choose one of these approaches there is a problem in fitting a distribution. Houston and Wall[55,p.106], who used the frequency-size approach, found that the five percent most frequent terms did not fit the log-normal distribution. Conversely, both Bennett[6] and Toma[115], who used the frequency-rank approach, discovered poor fits for the large rank, small frequency terms. In fact Toma had problems fitting both ends of the particular form of the Zipf distribution he was using. He says that the theoretical distribution "...leaves out from the actual distribution, 60 terms in the zone of the largest frequencies, while at the end of the lowest frequencies it lengthens appreciably the tail of the distribution." [115,p.268]. Booth[12] points out the problems of applying Zipf's original frequency-rank distribution to words of low frequency in the area of word frequencies in text.

The solution adopted here is to use the frequency-size analysis for the low frequency terms and the frequency-rank for the high frequency terms. This creates another problem. Where is the split between high and low frequencies to be? One of the reasons the frequency-rank analysis fails for low frequency terms is that, since the ranking is by frequency of occurrence,

after a certain point we get many terms with the same frequency and the ranks are not unique. Conversely, the frequency-size analysis fails because of the long tail of high frequency terms, with no terms for many of the frequencies. So the natural splitting point between high and low occurring terms is where ties in the ranks begin, from the rank point of view, or where gaps in the frequencies occur, from the size point of view. Looking at Tables 2,3,4, and 5, which show term frequencies for the four databases, the two points are close in most cases. In these tables, g(x) represents the number of terms occurring x times, f(r) represents the frequency of the rth ranking term, and F(r) the frequency of all terms of rank r or less.

4.5 DISTRIBUTIONS TESTED

A brief review of the theoretical distributions tested will now be given. Since most distributions investigated are discrete with no value for the zero case, the shifted discrete distribution functions are important. A shifted discrete distribution is one where the independent variable x is transformed to x+1, and the distribution now starts at one. The distributions reviewed are the shifted binomial, Poisson, negative binomial, shifted negative binomial, Zipf and log-rank.

A more detailed explanation can be found in Johnson and Kotz[59].

4.5.1 Shifted Binomial

The probability that the random variable takes on the value x for the shifted binomial is:

$$p(x)=\binom{N}{x-1} p^{x-1} q^{N-x+1} \quad x=1,2,\ldots N+1$$

where N, an integer, and $0<p<1$ are the parameters and $q=1-p$. The moment estimators of N and p are $p=1-v/m$ and $N=(m-1)/p$ where m is the sample mean and v is the sample variance.

4.5.2 Poisson

For the Poisson:

$$p(x)=e^{-m} l^{x}/x! \quad x=0,1,2,\ldots$$

The moment and maximum likelihood estimator for the parameter $l>0$ is the sample mean.

4.5.3 Negative Binomial

For the negative binomial:

$$p(x) = \frac{(n+x-1)!}{(n-1)!x!} \quad w^x (1-w)^n \quad x=0,1,2...$$

with parameters $0<w<1$ and $n>0$. The moment estimators are $w=1-m/v$ and $n=m(1-w)/w$.

For the shifted negative binomial:

$$p(x) = \frac{(n+x-2)!}{(n-1)!(x-1)!} \quad w^{x-1}(1-w)^n \quad x=1,2,3...$$

with parameters $0<w<1$ and $n>1$. The moment estimators are $w=1-(m-1)/v$ and $n=(m-1)(1/w-1)$. The maximum likelihood estimators must be solved using numerical iterative methods.

All three of these discrete distributions can be considered as part of the same family and to decide which one to use we consider the ratio of the variance to the mean (v/m). This ratio is less than one for the binomial, equal to one for the Poisson and greater than one for the negative binomial. This is a useful indicator to decide which of the three distributions to use, given that one of them is the correct distribution. See Johnston and Kotz[59] for a full explanation and an hypothesis test based on the variance-mean ratio.

4.5.4 Zipf

As noted is section 3.3.1, many versions of Zipf's distribution have been used. Two versions are used in this study, the simpler being

$$p(x) = k/x^b \qquad x=1,2,3,\ldots n$$

with parameter $b>1$ and $1/k = \sum_{x=1}^{n} 1/x^b$. Another version due to Mandelbrot[68], is also used

$$p(x) = a/(x+c)^b \qquad x=1,2,3,\ldots n$$

with parameters $b>0, c>0$, and $1/a = \sum_{x=1}^{n} 1/(x+c)^b$. This contains an additional 'shift' parameter 'c' which makes it more general. This will be referred to as the Mandelbrot-Zipf distribution. These Zipf distributions have been used for both the frequency-size and frequency rank approaches.

4.5.5 Log-rank

As implied by its name, the log-rank distribution is useful where the independent variable is the rank. If F(r) is the cumulative frequency distribution the most general form is

$$F(r) = a \log(r+c) + b \qquad r=1,2,3,\ldots n$$

with $c>0$, a and b parameters. This is just a generalization of the Bradford distribution which Asai[1] compares to versions by Brookes[16], Fairthorne[37] and

others for applications in journal citations. Asai uses a non-linear least squares method to estimate the parameters, and finds it fits citation data better than the other forms, but that it is still a poor fit for high rank, low frequency journals. A simpler log-rank distribution for the basic frequencies f(r) was also used:

$$f(r) = a \log(r) + b \qquad r=1,2,3,\ldots n$$

with parameters a and b. Since if r=1 then f(1)=b, we have a simple estimate for b, and it is also linear in a and b, we can use a linear least squares estimate for a. Of course we could use the full least squares estimates for both a and b.

4.6 FITTING THE TERM OCCURRENCE DATA

For the first three databases used, SPI, Cranfield and MEDLARS, the basic Zipf distribution provided the best fit for the low-frequency part and the simple log-rank function was best for the high-frequency portion (see Tables 6,7, and 8). So if E(x) is the expected number of terms with x occurrences, and E(r) is the expected number of occurrences of the rth ranked term, $E(x)=nk/x^b$ (n is the number of terms) for x=1,2,...H1 and E(r)=a log(r) + b for r=1,2,...H2. H1 is the maximum number of occurrences in the frequency-size part and H2

is the maximum rank in the frequency-rank part. This means the terms are split into two subsets, the H2 terms with a high number of occurrences ranked by the decreasing number of occurrences from one to H2, and the (n-H2) terms which are described by a frequency-size distribution. The maximum frequency which must be described by the frequency-size distribution is one less than the smallest frequency of occurrence in the frequency-rank part. So x must vary from one to H1=E(H2)-1. In the frequency-rank portion the independent variable is the rank r, and the dependent variable is the frequency of occurrence f(r). In the frequency-size portion different variables are needed, the frequency of occurrence, x, as the independent variable and the number of terms with x occurrences as the dependent variable g(x).

Other distributions tested at this time were:

1. frequency-rank Zipf for all terms
2. frequency-size Zipf for all terms
3. Brookes form of Bradford's law for cumulative rank
4. Negative binomial frequency distribution.

The first set of simulations were run using these distributions. For the CIJE database these distributions did not give a good fit. In particular, the low frequency portion of the terms was much lower than the

FIGURE 2

DISTRIBUTION OF INDEX TERMS

CIJE - MANDELBROT-ZIPF

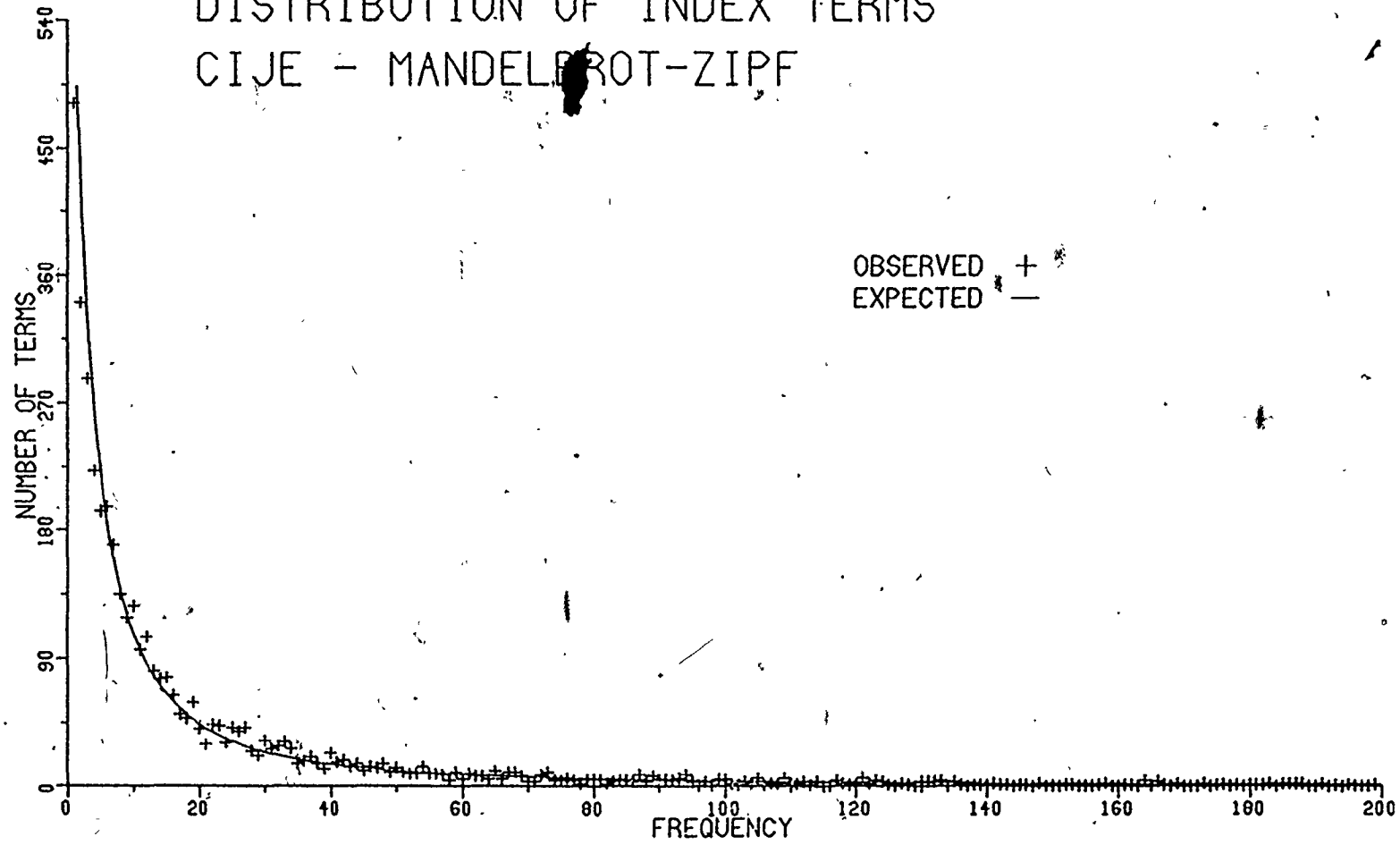
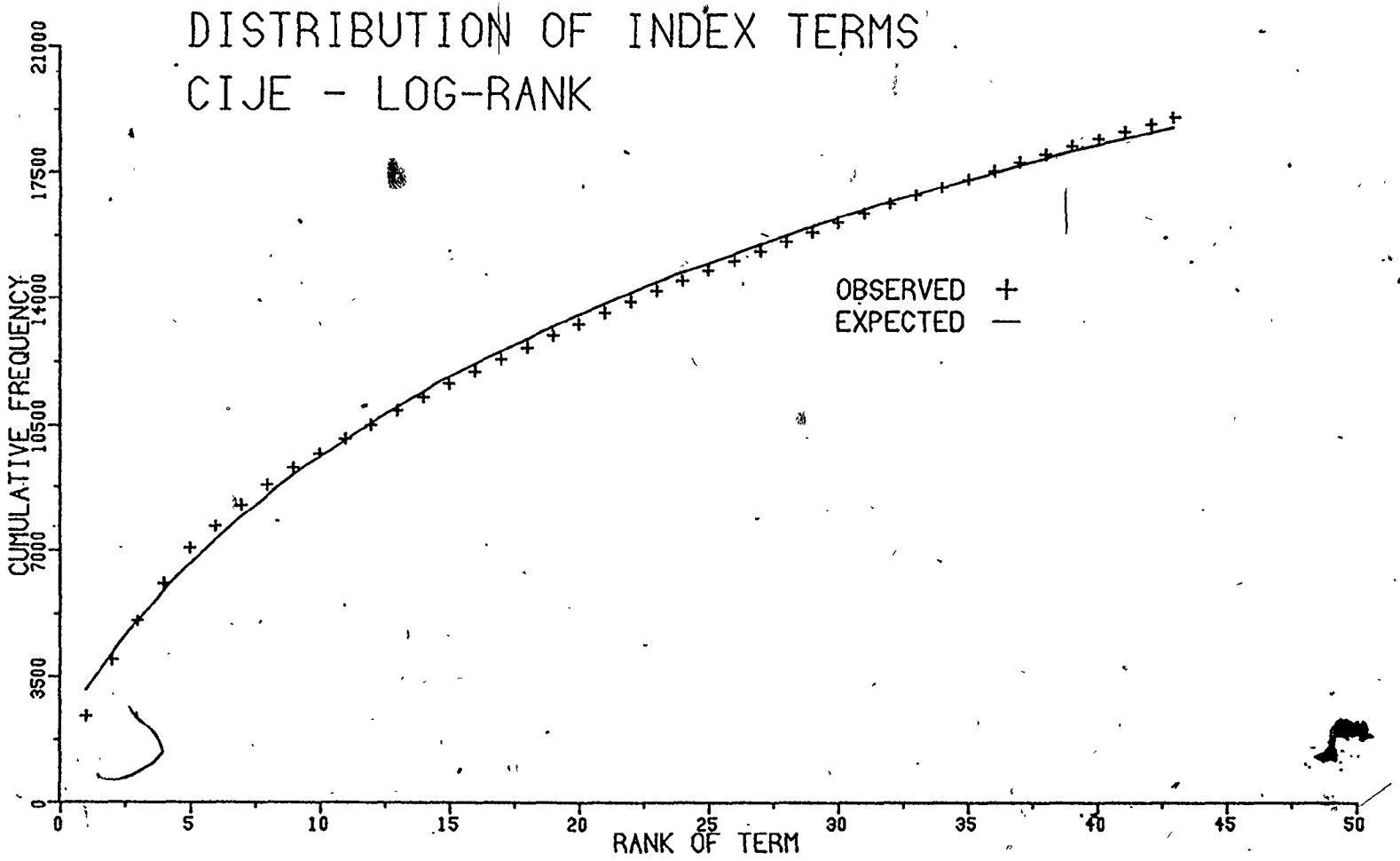


FIGURE 3

DISTRIBUTION OF INDEX TERMS

CIJE - LOG-RANK



Zipf distribution predicted, using the parameter estimation explained above. At this point the generalized Mandelbrot-Zipf distribution was tried, which was a much better fit (see Table 9A and Figure 2). The shift parameter 'c' seems to indicate the amount of control put on the vocabulary, the greater the value of c the more control is put on the vocabulary, and the fewer are the terms which occur only once or twice.

For the high frequency terms, a generalized log-rank was tried for the basic frequencies f(r) with little improvement, so the generalized log-rank was tried for the cumulative frequency distribution F(r) and a reasonably good fit was obtained (see Table 9B and Figure 3). For both of the log-rank distributions, the non-linear regression method was used to estimate the three parameters. There were no simple method of moments or maximum likelihood estimators, and the regression method does not require any assumptions about sampling from a population.

4.7 EXHAUSTIVITY OF INDEXING

By indexing exhaustivity is meant the number of terms used to index a document. In our databases there is a wide variation of average indexing exhaustivity,

ranging from 6.2 for SPI to 53.6 for Cranfield. Several simulation runs were made using the initial assumption that the distribution of indexing exhaustivity over documents was Poisson, from the main conclusion of Bird[7] on studying the distribution of indexing exhaustivity (which Bird calls the depth of indexing). On closer inspection, the variance-mean ratios, which must be one for the Poisson, were .705 for CIJE, .993 for SPI, 8.271 for MEDLARS, and 10.339 for Cranfield. If we look closely at Bird's work we see he also tried negative binomial and binomial distributions and found they fit in certain cases. From the evidence of the variance-mean ratio, we should use the binomial for CIJE, Poisson for SPI, and negative-binomial for MEDLARS and Cranfield.

In fact the shifted negative binomial and shifted binomial distributions fit reasonably well as is shown in Tables 10 to 12 and Figures 4 to 6 (complete data was unavailable for SPI). This assumes that one of these three forms is onappropriate for the distribution of indexing exhaustivity. Bird gives a stochastic argument for this case, saying that these three distributions are the natural ones to consider. The variance-mean ratio shows that CIJE was very regular and controlled, with a maximum of 21 index terms assigned to one document, while at the other extreme the Cranfield database has a high

FIGURE 4

DISTRIBUTION OF EXHAUSTIVITY

MEDLARS - SHIFTED NEGATIVE BINOMIAL

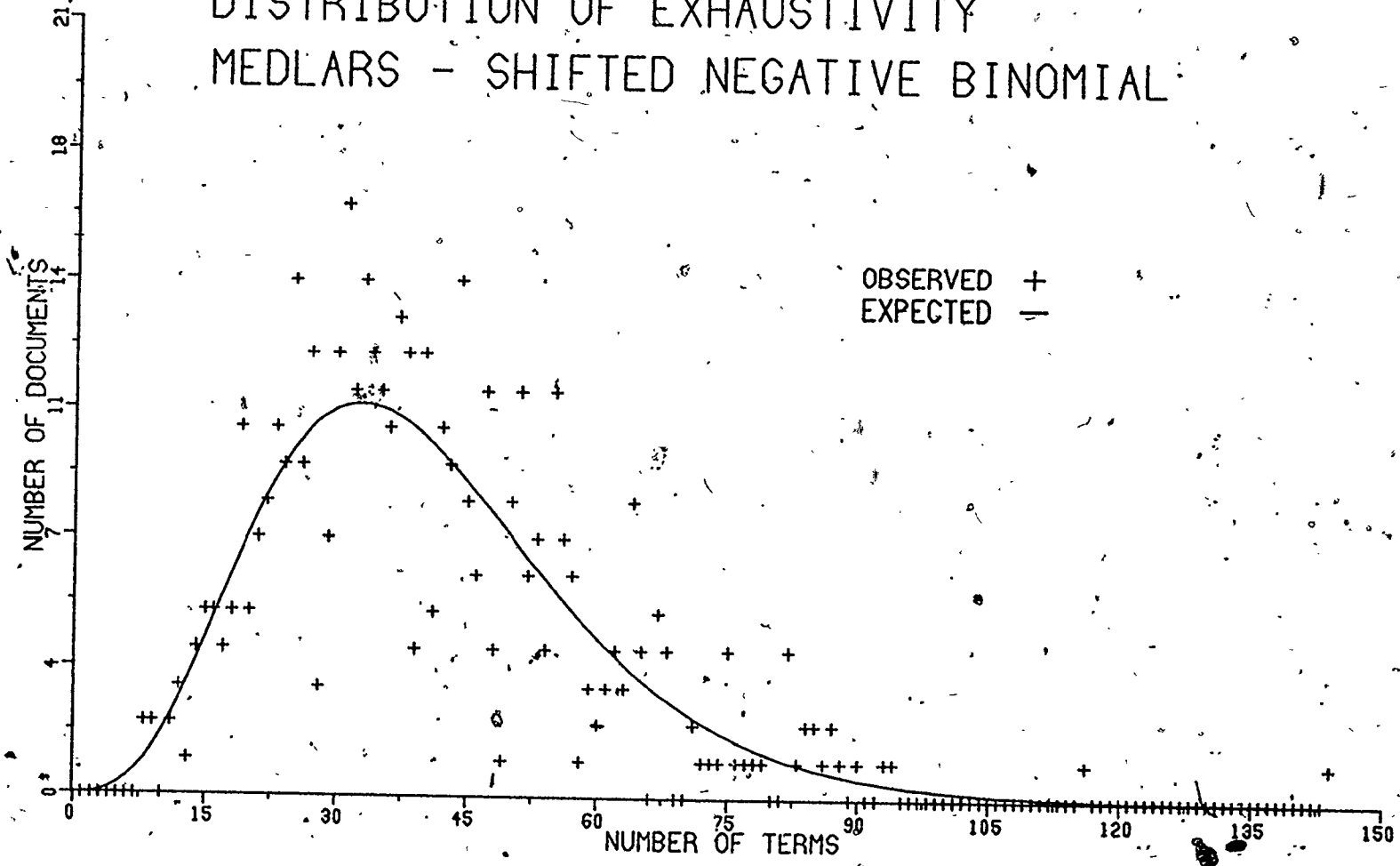


FIGURE 5

DISTRIBUTION OF EXHAUSTIVITY

CRANFIELD - SHIFTED NEGATIVE BINOMIAL

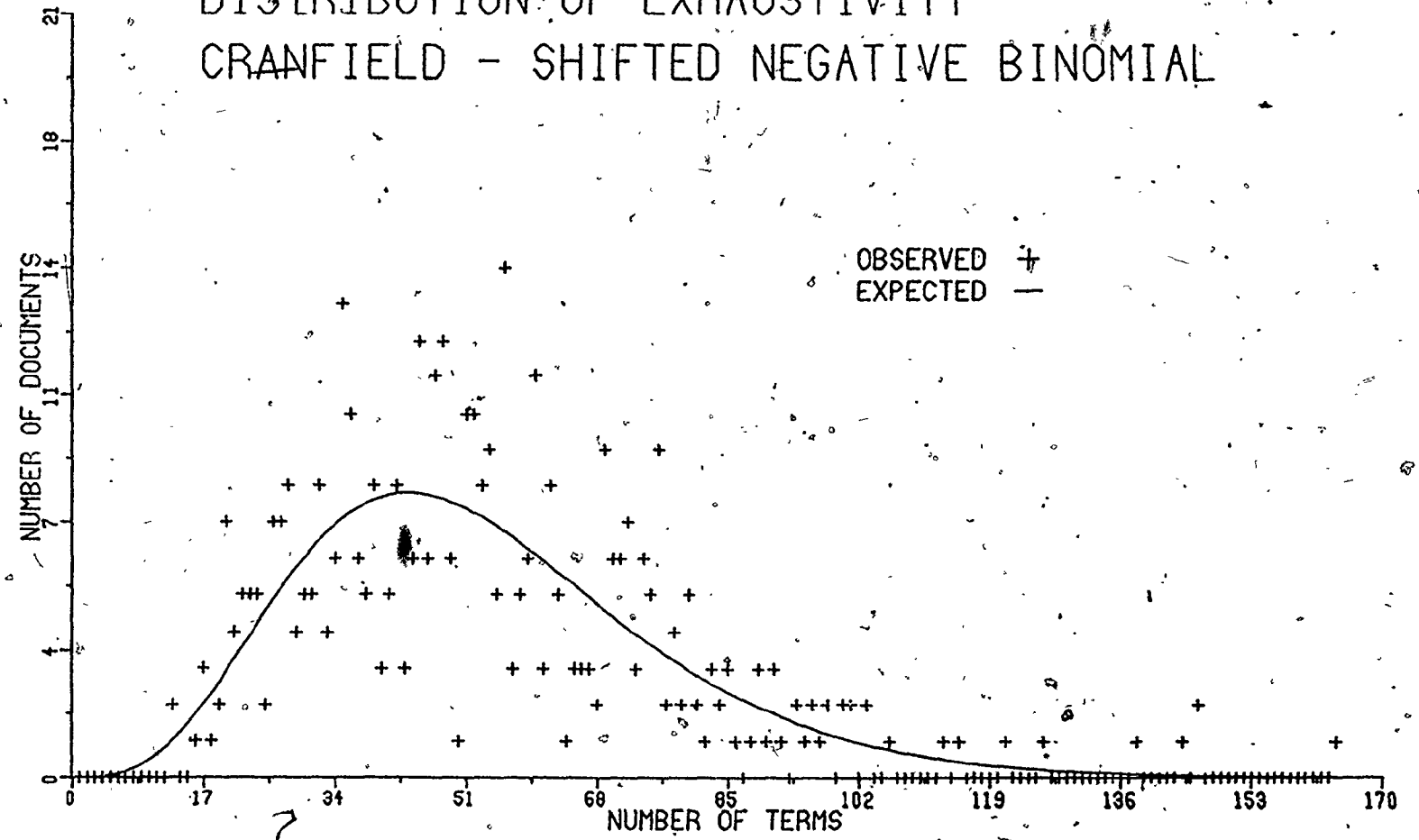
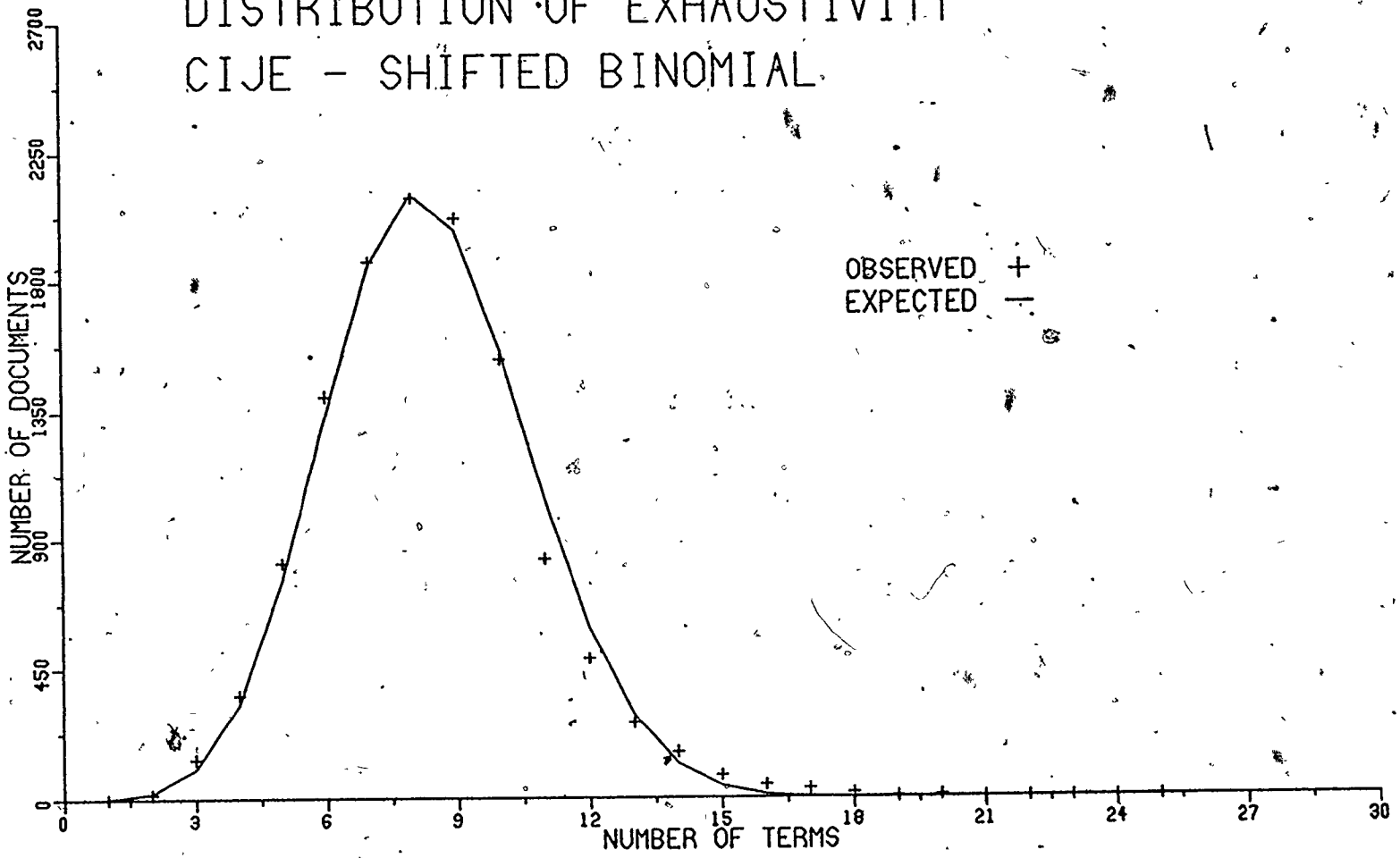


FIGURE 6
DISTRIBUTION OF EXHAUSTIVITY
CIJE - SHIFTED BINOMIAL



variation in the number of terms per document with up to 144 terms assigned to one document.

4.8 EXHAUSTIVITY OF INDEXING IN QUERIES

As we can see from Table 13 the query statistics show that for the twenty-four Cranfield queries the Poisson distribution of indexing exhaustivity is most likely with a variance-mean ratio of 0.98. For the MEDLARS, with a variance-mean ratio of 2.435, the negative binomial distribution is a likely candidate. Since we have a fairly small sample of twenty-four queries in each case, no goodness-of-fit tests were attempted. More data is needed for queries to make firm conclusions. Consequently, the Poisson distribution of the number of terms per query was used in the simulation model.

4.9 RELEVANCE DISTRIBUTIONS

For the Cranfield and MEDLARS databases there were twenty-four queries with relevance judgements for all the documents, i.e. relevant or nonrelevant. In order to simulate the function a: Q -> S(D) we used the distribution of the number of matching query terms over relevant documents, and over nonrelevant documents. As

we can see from Tables 14 and 15, these have the form of negative binomial distributions. The parameters n and w will , in general, have different values for the relevant and nonrelevant sets. If p(t) is the probability of t query terms matching a relevant (or nonrelevant) document, we have:

$$p(t) = \frac{(n+t-1)!}{(n-1)!\ t!}\ w^t\ (1-w)^n \qquad t=0,1,2,3...$$

4.10 DISTRIBUTIONS OF CO-OCCURRENCES

In order to use term dependencies in the final model, we will need the conditional distribution of the number of co-occurrences with a term given its occurrence frequency. The number of co-occurrences of a term with another term is the number of documents that contain both terms. Some of the data from the SPI database suggested this was a Zipf frequency size distribution. The parameter appeared to be a linear function of the logarithm of the number of occurrences y. Thus, the probability of x co-occurrences with another term, given that a term occurs y times is

$$p(x/y) = \begin{cases} k_y/(1+x)^{b_y} & x=0,1,2,...y \\ 0 & \text{for } x>y \end{cases}$$

where $b_y = c_0 + c_1 \log y$

$$k_y^{-1} = \sum_{x=0}^{y} [1/(1+x)^{b_y}]$$

The parameters which need to be estimated are c_0 and c_1. To estimate c_0, note that the expected number of terms which co-occur with a term which occurs once is just a-1, where a is the average number of terms per document. Then

$$n\ p(1/1) = a-1$$

$$n\ k_1 / 2^{c_0} = a-1$$

$$n\ /(2^{c_0}+1) = a-1 \quad \text{since } k_1 = 2^{c_0}/(2^{c_0}+1) \text{ from above}$$

$$2^{c_0} = (n-a+1)/(a-1)$$

$$c_0 = \log[(n-a+1)/(a-1)]/\log 2$$

To estimate c_1, we will assume that when y is n_1, the frequency of the most frequent term, that the distribution of p(x/y) has the same parameter as the original Zipf frequency-size distribution with parameter b,

$$b_n = b = c_0 + c_1 \log n_1$$

$$c_1 = (b - c_0)/\log n_1$$

However, the expense and time for a complete goodness-of-fit test was impractical and a complete validation of this assumption has not been made. If the co-occurrence data in the simulated document set agrees with the observed co-occurrence data, the assumption will be considered validated.

CHAPTER 5

THE SIMULATION MODEL AND RESULTS

5.1 SIMULATION AND THE GENERAL MODEL

The general model from Chapter 2 must be made more specific in order to apply it to simulating databases. Some of this specification has occurred in the development of the distributions in the previous chapter. All this information will now be put together into a specific model which can be parameterized and implemented in a simulation program. Simulations will then be carried out for specific values of the parameters and validated with real data.

5.2 THE SPECIFIC MODEL

Remembering from Chapter 2 that a bibliographic retrieval system can be represented as a tuple, we will now concentrate on the logical elements of the system, and not the physical. The general model was:

$$B = \langle D, Q, T, P, A, i, r=r_1\ r_2\ r_3\ r_4, a, b \rangle$$

The individual functions used in the specific model are:

i: D -> S(T); document indexing is the assignment of subsets of 'keywords' from the set of terms or thesaurus.

r_1 : Q -> S(T); queries are also indexed by subsets of terms.

r_2 : S(T) -> A; for the Cranfield and Medlars databases, the retrieval algorithm involves the calculation of the cosine similarity measure between documents and queries.

r_3 : A -> V(P); the algorithm scans the document storage locations sequentially to match against the query.

r_4 : V(P) -> O(D); the output of the retrieval is an ordering of the documents based on the cosine correlation similarity measure.

a:Q -> C_2(D); assessment involves assigning each document to a relevant or nonrelevant subset.

b: O(D) x C_2(D) -> F(R);, evaluation involves determining the recall-precision curve. The curve is determined by microaveraging recall and precision at each rank, where document ranks are based on the value of the document-query cosine measure.

This is still not specific enough to form an algorithm for implementing a simulation model. The basic outline of

the simulation process is:

1. Generate the document set D and all term assignments (function i).

2. Generate the queries and their term assignments (function r_1).

3. For each query-document pair calculate cosine correlation (function r_4) and decide if the document is relevant or nonrelevant to that query (function a).

4. Calculate the average recall and precision at each rank over all queries (function b).

5.3 GENERATING THE DOCUMENT SET AND TERM ASSIGNMENTS

In order to assign terms to documents, we will first use the distribution of indexing depth to decide how many terms a document will be assigned. If we assume that index terms are assigned independently, we can easily assign the index terms one at a time by sampling from the basic distribution of index terms. Since we want to compare this assumption to the assumption that term dependence is an important factor in the performance of retrieval systems, we must have a method of incorporating term dependencies into the term assignments.

5.3.1 Term Probability Functions

The functions used to accomplish this are as follows:

g(r): the probability that the first term assigned to a document is the rth term. g(r) is derived from the basic split frequency-size, frequency-rank distribution for the occurrence of terms.

g(r/s): the probability that a subsequent term assigned to a document is the rth term, given that the sth term is the first assigned. This can be derived from the distribution of co-occurrences as follows. From section 4.10 for a term which occurs y times, the probability of x co-occurrences with another term is:

$$p(x/y) = \begin{cases} k_y/(1+x)^{b_y} & x=0,1,2,\ldots y \\ 0 & \text{for } x>y \end{cases}$$

where $b_y = c_0 + c_1 \log y - 1$

$$k_y^{-1} = \sum_{x=0}^{y} [1/(1+x)^{b_y}]$$

$c_0 = [\log(n-a+1)/(a-1)]/\log 2$

$c_1 = (b-c_0)/\log n_1$

n = number of terms

a = average number of terms per document

n_1 = frequency of most frequent term.

So let x be the number of co-occurrences with other terms, y be the number of occurrences of the second and subsequent terms, and z be the number of occurrences of the

first term. Assume $p(x/y)$ is independent of z and that $p(x/z)$ is independent of y, in other words the number of co-occurrences does not depend on any of the other terms assigned, only the two under consideration.

Then Bayes theorem gives:

$$p(y/x) = p(x/y)\ p(y)/p(x) = p(x/y)\ p(y)/\sum_{v=1}^{n_1} p(x,v)$$

$$= p(x/y)\ p(y)/\sum_{v=1}^{n_1} p(x/v)p(v)$$

and the probability of selecting a term which occurs y times, given the first term occurs z times :

$$p(y/z) = \sum_{x=0}^{h} p(x,y/z) \qquad h = \min(y,z)$$

$$= \sum_{x=0}^{h} p(y/x)p(x/z)$$

$$= \sum_{x=0}^{h} p(x/z)p(x/y)p(y)/\sum_{v=1}^{n_1} p(x/v)p(v)$$

$$= \sum_{x=0}^{h} k_z k_y k/(1+x)^{b_z+b_y}\ y^b\ /\ \sum_{v=1}^{n_1} p(x/v)p(v) \ .$$

To get $g(r/s)$ the function $p(y/z)$ is transformed from a frequency function to a rank function. Remember that e is the expected frequency of the ith ranked term. For r not equal to s:

$$g(r/s) = \begin{cases} \sum_{v=e_{r+1}+1}^{e_r} p(v/e_s) & r < H2 \\ p(e_r/e_s)\ (e_r+c)^b\ /k(n-H2) & r > H2 \end{cases}$$

5.3.2 The Term Assignment Procedure

The procedure for assigning terms using second order dependence information is to assign the first term from the basic distribution of occurrences g(r), then to use g(r/s) to assign the remaining terms, where the sth term is the first assigned. The assumption is that once the first term has been selected it essentially determines the subject area of the document, an hypothesis not inconsistent with indexing practice.

5.3.3 Example Of Term Probabilities

An example of the calculation of g(r) and g(r/s) is now presented. Suppose there are only five terms whose basic occurrence distribution is a split distribution; log-rank for the frequency-rank portion and a Mandelbrot-Zipf for the frequency-size portion. The parameters are n=5, H2=2 (number of terms in frequency-rank), a=-3 b=8 for the log-rank distribution, and b=2 c=0 for the frequency-size portion. Let e_i be the expected frequency of the ith ranked term. First calculate the frequency rank portion:

$$e_i = 8 - 3\log(i) \quad \text{for } i=1,2$$

$e_1 = 8$ $e_2 = 8 - 3 \log 2 = 5.9 = 6$ (rounded)

Since the smallest frequency of a term in the frequency rank portion is six, the maximum frequency of any term in the

frequency-size portion is five, and H1=5. To calculate the "a" parameter for the Mandelbrot-Zipf distribution (see 4.5.4):

$$1/a = \sum_{x=1}^{5} 1/x^2 = 1.56 \quad \text{so } a = 0.68 .$$

For the frequency-size portion, $p(x)=0.68/x^2$ where $p(x)$, is the probability that a term will occur x times. Let $E(x)$ be the expected number of terms which occur x times, then $E(x)=(n-H2)p(x)$ and $E(1)=(5-2)(0.68)=2.04$, $E(2)=0.51$, $E(3)=0.23$, $E(4)=0.13$, and $E(5)=0.08$. If we round to the nearest integer, since these represent numbers of terms, we are left with $E(1)=2$ and $E(2)=1$.

We now convert the split distribution into a rank probability distribution $g(r)$. The third ranking term, ranked by number of postings, occurs twice so $e_3=2$. There are two terms occurring once which will have ranks four and five, so $e_4=1$ and $e_5=1$. Note that the total number of postings is just the sum of e_i, namely 18. We have directly that:

$g(1)= 8/18 = 0.45$ $\quad$ $g(2)= 6/18 = 0.32$

$g(3)= 2/18 = 0.11$ $\quad$ $g(4)= g(5)=1/18 = 0.06$

These are the simple probabilities of occurrence of the five terms.

To incorporate the dependence information, $p(y/z)$ must be calculated. First calculate $p(x/y) = k_y/(1+x)^{b_y}$ for

y=1,2,...8 following the notation in the previous section (5.3.1). If the average number of terms per document is three then:

$c_0 = [\log(5-3+1)/(2-1)]/\log 2 = \log 3/\log 2 = 1.58$

$c_1 = (b-c_0)/\log n_1 = (1-1.58)/\log 8 = .20$

We can now calculate the parameters for the distributions of co-occurrences:

$b_1 = 1.58,\ b_2 = 1.72,\ b_3 = 1.80,\ b_4 = 1.86,$

$b_5 = 1.90,\ b_6 = 1.94,\ b_7 = 1.97,\ b_8 = 2.00;$

$k_1 = 0.75,\ k_2 = 0.69,\ k_3 = 0.66,\ k_4 = 0.65,$

$k_5 = 0.65,\ k_6 = 0.65,\ k_7 = 0.65,\ k_8 = 0.65.$

Next p(x), the probability of x co-occurrences, is calculated:

$$p(x) = \sum_{v=1}^{n_1} p(x/v)\ p(v)$$

$$= \sum_{v=x}^{n_1} k_v/(1+x)^{b_v}\ \ (k/v^b)$$

$$= \sum_{v=x}^{8} (k_v/(1+x)^{b_v})\ (0.68/v^2)$$

so $p(0) = 0.75,\ p(1) = 0.24,\ p(2) = 0.034,\ p(3) = .009,$

$p(4) = .003,\ p(5) = .001,\ p(6) = .0006,\ p(7) = .0002,$

$p(8) = .00008.$

Now calculate p(y/z), the probability of a second or subsequent term having y occurrences given that the first term selected has z occurrences:

$$p(1/8) = \sum_{x=0}^{1} k_8 k_1 k/(1+x)^{b_1+b_8}\ y^2 p(x)$$

$$= \sum_{x=0}^{1} (0.65)(0.75)(0.68)/(1+x)^2\ p(x)$$

$$= 0.33/(0.75) + 0.33/(4)(0.24)$$

$$= 0.56$$

Similarly, p(2/8)= 0.16, p(3/8)=.086, p(4/8)=.057,
p(5/8)= .042, p(6/8)= .034, p(7/8)= .030, p(8/8)= .030
p(1/1)= .687, p(2/1)= .154, p(3/1)= .065, p(4/1)= .036,
p(5/1)= .023, p(6/1)= .016, p(7/1)= .011, p(8/1)= .009.
The other probabilities for y=1 to 8 and z=1 to 8 are calculated in the same way.

Finally we can get g(r/s), the probability that the rth term is assigned given that the sth term is the first assigned:

$$g(2/1) = \sum_{v=e_{r+1}}^{e_r} p(v/e_s) \quad \text{since } r=$$

$$= \sum_{v=3}^{8} p(v/8)$$

$$= p(3/8) + p(4/8) + p(5/8) + p(6/8)$$

$$= 0.085 + 0.057 + 0.042 + 0.035$$

$$= 0.219$$

$$g(3/1) = p(e_r/e_s)\ (e_r)\ /k(n-H2) \quad \text{since } r>H2$$

$$= p(2/8)\ (4)/0.68(5-2)$$

$$= 0.319$$

Similarly, g(4/1)= .273, g(5/1)= .273,
g(1/2)= .045, g(3/2)= .326, g(4/2)= .274, g(5/2)= .274,
g(1/3)= .024, g(2/3)= .169, g(4/3)= .301, g(5/3)= .301,
g(1/4)= .020, g(2/4)= .139, g(3/4)= .302, g(5/4)= .337,
g(1/5)= .020, g(2/5)= .139, g(3/5)= .302, g(4/5)= .337.

5.4 SIMULATING QUERIES AND RELEVANCE JUDGEMENTS

The queries are simulated in exactly the same way as documents, except that the parameters for the depth of indexing may be different.

To simulate relevance judgements for the Cranfield and Medlars databases, we use the distribution of the number of query terms over relevant and nonrelevant documents, which was negative binomial. Let u=1 if a document is relevant to a query and u=0 if nonrelevant. Let t be the number of query terms contained in the document, let the probability a document contains t query terms, given that it is relevant, have a negative binomial distribution as in Chapter 4, with parameters n_1 and w_1, and let the probability a document contains t query terms given that it is nonrelevant have a negative binomial distribution with parameters n_0 and w_0. Then by Bayes' theorem:

$$p(u/t) = \frac{p(t/n_u, w_u)\ p(u)}{p(t/n_0, n_0)\ p(0) + p(t/n_1, n_1)\ p(1)}$$

The unconditional probability p(u=1) is the over-all probability of relevance and p(u=0)=.1 - p(u=1). Using this probability function, a 0 or 1 relevance value can be randomly generated for each document, given a query.

The parameters n_i, w_i, i=0,1 are estimated from the sample mean m and standard deviation s of the number of query terms per document and the number of documents

containing zero terms z in sets of relevant and nonrelevant documents from the collections being simulated. The estimators are:

$$w_i = 1 - m_i / s_i$$

$$n_i = \text{int}\,[\log z_i]/\log(1-w_i) \qquad i=0,1$$

The probability of relevance p(u=1) is estimated from the test collection as:

$$p(u=1) = \frac{\text{(average number of relevant documents per query)}}{\text{(total number of documents)}} = \frac{q}{m}$$

5.5 THE SIMULATION PROGRAMS

The model has been implemented as a Fortran program which is run on the University of Western Ontario's CYBER-170 computer. Details can be found in Appendix 1. There are four versions of the simulation program. The first two versions 1a and 1b assume a Poisson distribution of indexing depth, and a split distribution for the distribution of terms -- Zipf for the frequency-size, simple log-rank for the frequency-rank.

The basic parameters used as input to the program are:

M - the number of documents

N - the number of terms

Q - the number of queries

H2 - the number of terms generated by the log-rank distribution

H1 - the maximum frequency generated by the Zipf frequency distribution

MAXFRQ - the number of postings of the most frequent term

N2 - the number of postings of the second most frequent term

FRACTN - $\sum_{r=1}^{H2} f(r)\log r$, where $f(r)$ is the frequency of the rth ranked term

A - the average number of terms per document

B - the average number of terms per query

G - the average number of relevant documents per query (generality)

B1 - the parameter for the Zipf frequency distribution part of the distribution of index terms.

Given these basic parameters as input, we must estimate parameters for the various distributions. A gives us the parameter for the Poisson distribution of indexing exhaustivity, and similarly B is the Poisson parameter for the distribution of query indexing exhaustivity. For the

frequency-rank portion of indexing exhaustivity,

$$f(r) = a \log(r) + b \quad r=1,2,3,\ldots H2$$

where b is estimated by MAXFRQ and a is estimated by least squares as:

$$a = (f(1) \sum_{r=2}^{H2} \log r - \sum_{r=2}^{H2} f(r) \log r) / \sum_{r=2}^{H2} (\log r)$$

$$a = (MAXFRQ \sum_{r=2}^{H2} \log r - FRACTN) / \sum_{r=2}^{H2} (\log r)$$

For the Zipf frequency portion we have the parameter B1 input directly. It could also be estimated from :

$$B1 = \log(MAXFRQ/N2) / \log 2 .$$

Version 1a assumes the independence of index terms and version 1b assumes the second order dependence model explained in the previous section, where the parameters for the co-occurrence distributions are estimated from the basic input parameters.

5.6 VERSIONS 2A AND 2B FOR CIJE

For the CIJE database the distributions used in version 1 did not fit very well, so a new version was written, version 2a for CIJE with the independence assumption, and version 2b for CIJE with the dependence model. These two versions assume a binomial distribution of indexing exhaustivity. For the distribution of terms, the frequency-rank part was a generalized Bradford for the

cumulative F(r) and for the frequency size portions the Mandelbrot-Zipf frequency distribution is assumed. The implementation of the co-occurrence distributions remains the same.

The new parameters needed for these versions are:

V - the variance of the number of terms per document. The parameters of the binomial distribution of indexing exhaustivity can then be estimated by the method of moments from A and V.

A1,B1,C1, - the three parameters for the Mandelbrot-Zipf distribution

$$g(x) \triangleq A1/(x+C1)^{B1}$$

A2,B2,C2 - the three parameters for the generalized Bradford in relative frequency form:

$$F(r)/F(H2) = A1 \log(r+C1) + B1$$

The parameters FRACTN and B1, and any query parameters from versions 1a and 1b are not used.

A disadvantage of this approach is that the six parameters needed for the term distributions are not simple statistics from the sample database, but require an additional program to estimate them prior to the simulation run.

5.7 TESTING THE VALIDITY

There are several ways of analyzing the output from the simulations to determine validity. Because of the exploratory nature of this research, only replicative validity and to a small extent predictive validity has been tested. For a comprehensive predictive validation, more data is needed, especially for queries and evaluations, and for structural validity, a much greater understanding of the mechanisms of a retrieval system is needed. The following six methods are used:

1. Comparing the theoretical distributions of term frequencies and indexing exhaustivity with the actual numbers in the databases. These are exactly the tests carried out to develop the model in Sections 4.6 and 4.7.

2. Predicting the distributions of term occurrences and indexing exhaustivity from the CIJE-2 database by using the parameters estimated in CIJE.

3. Comparing the distributions of the number of query terms in relevant and non-relevant documents in the real set and the simulated set using chi-square.

4. Comparing by eye the recall-precision curves for the real and simulated sets, since there is no standard statistical test for the significant difference of two

recall-precision curves.

5. Applying a two-sample Kolmogorov-Smirnov test to compare the recall versus cosine similarity measure distribution of real and simulated data.

6. Comparing the real and simulated co-occurrence distributions using chi-square.

5.7.1 Prediction Of CIJE-2 Distributions

The distribution of the term occurrences and indexing exhaustivity for CIJE-2 were compared to the theoretical distributions with parameters estimated from the CIJE database (Tables 28 and 29). The distribution of index terms in CIJE fit the theoretical distributions slightly better than the CIJE-2 term distributions as measured by chi-square. The distribution of exhaustivity actually gave a slightly better fit for CIJE-2 than for CIJE. This shows that the characterizations of the CIJE database can at least be generalized to a larger database from Current Index to Journals in Education. This is a simple test of predictive validity, considering the CIJE database as a sample from the total database of articles from Current Index to Journals in Education.

FIGURE 7

RECALL-PRECISION MEDLARS

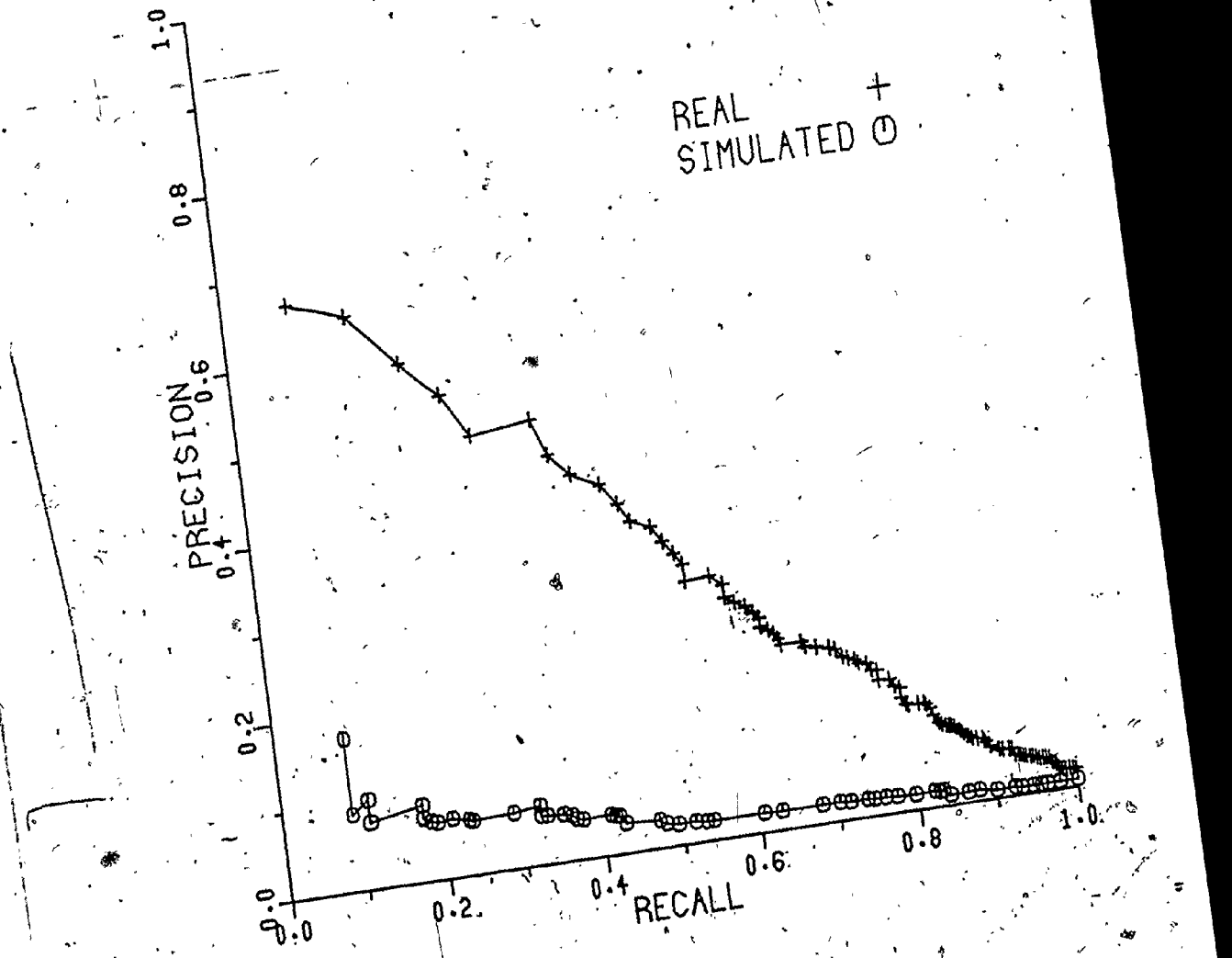
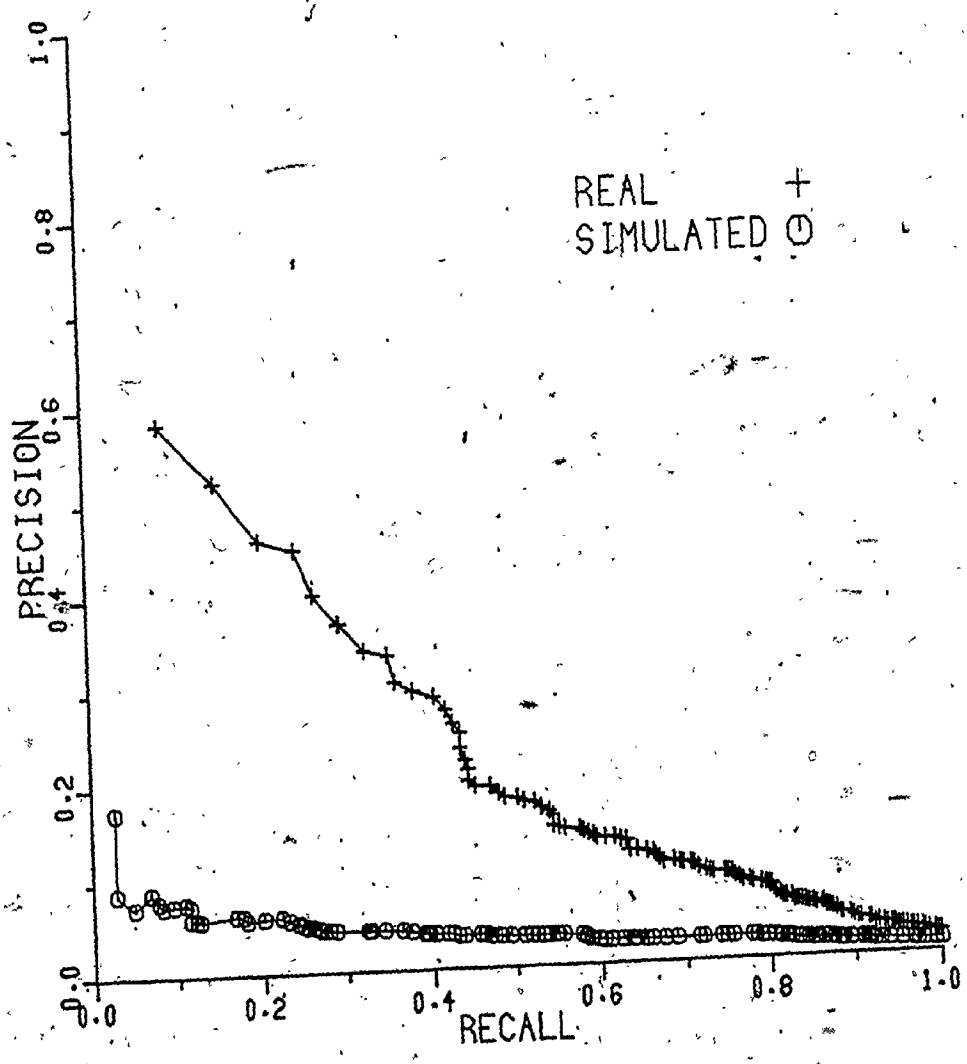


FIGURE 8

RECALL-PRECISION CRANFIELD



5.7.2 Recall-Precision Curves

For the Cranfield and Medlars databases we can compare the recall-precision curves of the real and the simulated databases (see figures 7 and 8). These curves were obtained by microaveraging recall and precision at each rank, where the ranks are determined by the cosine similarity between document and query. There is obviously a large discrepency between the real and simulated sets.

5.7.3 Recall Hypothesis Test

Also, recall can be considered, in some sense, as a cumulative distribution function, since it is the proportion of the relevant documents retrieved for each cosine value. Thus, the Komogorov-Smirnov two sample test may be used to test the hypothesis that the two empirical cumulative distribution functions appear to come from the same underlying population function. The test statistic is:

$$D = \sup_x [F_1(x) - F_2(x)]$$

where x represents the cosine measure and $F_1(x)$ and $F_2(x)$ represent the corresponding recall values for the real and simulated systems. For a significance level α, H_o the hypothesis of no difference is rejected for :

$$D > [-1/2(1/k + 1/l) \log(\alpha/2)]^{\frac{1}{2}}$$

where k and l are the number of points (i.e. unique cosine

measure values) in the real and simulated distributions respectively. The results of the tests for a significance level of .05 are given in Table 16. This shows that although, from the earlier tests, we are simulating the marginal term distribution validly, in the area of the query matching the documents the model is not working.

5.7.4 Number Of Query Terms In Documents

To investigate this further, we look at the distribution of the number of query terms matching in real and simulated sets in Tables 17 and 18. In both Cranfield and Medlars, the simulated set contains a larger number of documents with zero and one matching terms and a smaller number with a higher number of matching terms. Since documents with a higher number of matching terms are more likely to be relevant than not, the net effect is to reduce the number of relevant documents in the simulated as opposed to the real databases. This data in conjunction with that in Tables 14 and 15, indicates that our limited second-order dependency model is inadequate to represent the full dependency structure of the index terms. Dependencies of order higher than two seem to be occurring.

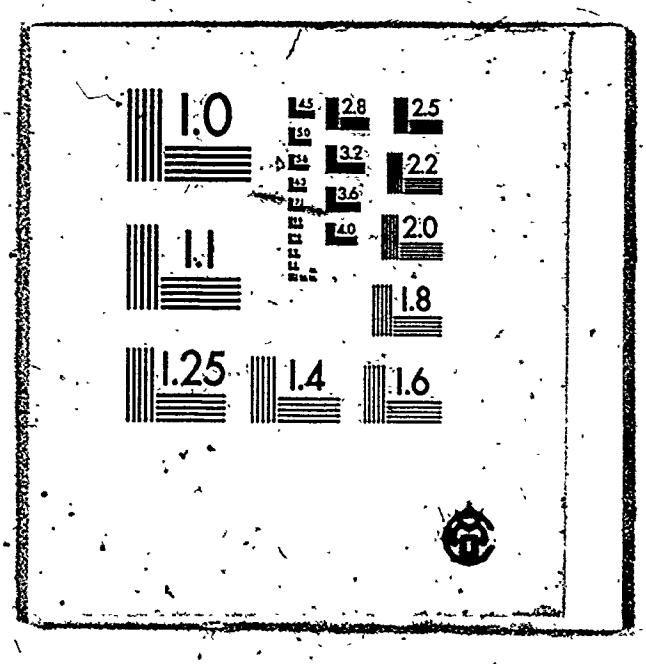
5.7.5 Comparison Of Co-occurrence Distributions

For the Medlars and Cranfield databases, complete co-occurrence distributions were calculated for both the real and simulated data, that is the number of occurrences of each unique pair of terms was calculated. This distribution was not simulated directly, as was the simple occurrence of terms, so it is some indication that the second order term dependence is working. The distributions were calculated for the original 'real' databases, the term independence model, and the term dependence model. The results are presented in Tables 19 and 20.

This discrepency is one area of the simulation which needs more investigation. At first it was thought that the Poisson depth of indexing was not assigning enough documents with a large number of terms and therefore fewer pairs would result. But a simulation with the negative binomial distribution of indexing depth (version 1c) in fact slightly reduced the number of unique pairs and total number of pair occurrences. Also the recall precision curve was very close to version 1b with the Poisson distribution of indexing depth. This shows that moderate changes in the depth of indexing distribution have very little effect on the final co-occurrence distribution.

The simulated independent model co-occurrences drop off much more quickly than either the real or the simulated dependent model; and the dependent model is reasonably good except for the tail, which contains pairs which occur a large number of times. It is conjectured that a split distribution of term co-occurrences, as is used for term occurrences, would provide a better fit. If this were used for the conditional probability of x co-occurrences given a term occurs y times, we might get a better fit. However, this approach has not as yet been implemented, because of difficulties in parameter estimation.

For the CIJE database with 12167 documents, a full generation of all pairs would be costly, so a random sample of 3000 pairs from each of the real, simulated independent and simulated dependent sets was taken. This was accomplished by randomly sampling pairs of terms from the inverted index, until 3000 co-occurring pairs were obtained. The results are in table 22. The chi-square test shows that the dependent model is a definite improvement on the independent model, but there is room for even more improvement.



5.8 CLUSTERING TERMS

A second approach to incorporating term dependence in the document set model was also investigated. The basic idea is to cluster terms which are dependent and to treat the clusters as single entities for the purposes of simulation. Each cluster is used as a 'hyper-term' and represents the Boolean combination of all the terms in the cluster 'OR'ed together. If there is a set of term clusters which is a partition, and for which the inter-cluster similarity is very small, this would be a good approximation to an independent set of clusters. When the documents are considered as cases and the terms as variables, the cluster interpretation of Tyron and Bailey[115,p118] applies. They suggest that the clustering of terms (variables) is a discrete form of factor analysis, thus the clusters or hyper-terms represent discrete factors.

The data base used to test this method was a subset of the SPI set, consisting of 242 documents and 715 terms. In order to make the computations easier only the 239 terms which occurred more than once were used in the clustering. The terms which occurred once could only affect the clustering slightly, as they would not have high correlations with frequently occurring terms. The similarity coefficient between

4. Stop the clustering when the similarity between clusters is no longer greater than zero.

5. Replace all term occurrences by their cluster number, thereby creating 'hyper-terms'.

6. Repeat.

After the first iteration there were seventy-four clusters from the original 239 terms. The distribution of all inter-cluster correlations is in Table 24. There are a large number of very small negative correlations, only one pair of clusters has the highest correlation of 0.5, and only a relatively small number of correlations between 0.2 and 0.5. This is a definite improvement over the original distribution of inter-term correlations (Table 23).

After the second iteration (Table 25), with twenty-seven clusters the highest correlation between clusters is 0.31 with only fifteen correlations (4.3%) greater than 0.1 . Clearly, the iterative procedure is reducing the hyper-term correlation, but at the expense of reducing the identity of the 239 terms to twenty-seven hyper-terms.

At this point, further iteration did not seem feasible, but in general it is difficult to give a

procedure which will decide on the stopping point, since the correlations will never be all zero.

The next problem to be explored is the problem of simulating the hyper-terms. The simplest model for such a simulation would assume that the size of a cluster and the frequencies of the terms within the cluster were independent. It was therefore hypothesized that there was a zero correlation between the size of the cluster and the average term frequency of terms in the cluster. The product-moment correlation coefficient for these variables was calculated (Table 26) and found to be 0.59 and the null hypothesis of zero correlation was rejected. The distribution of the size of the clusters (Table 27) does not give any indication of a regular distribution either.

There is much more to be done in investigating which terms to include initially, which type of similarity coefficient and clustering algorithm to use, and when to stop the hierarchial clustering algorithm. Most of all, a method of using the clusters to simulate the document set must be implemented. At this stage, the clustering provides further evidence of the inadequacy of the term independence model.

procedure which will decide on the stopping point, since the correlations will never be all zero.

The next problem to be explored is the problem of simulating the hyper-terms. The simplest model for such a simulation would assume that the size of a cluster and the frequencies of the terms within the cluster were independent. It was therefore hypothesized that there was a zero correlation between the size of the cluster and the average term frequency of terms in the cluster. The product-moment correlation coefficient for these variables was calculated (Table 26) and found to be 0.59 and the null hypothesis of zero correlation was rejected. The distribution of the size of the clusters (Table 27) does not give any indication of a regular distribution either.

There is much more to be done in investigating which terms to include initially, which type of similarity coefficient and clustering algorithm to use, and when to stop the hierarchial clustering algorithm. Most of all, a method of using the clusters to simulate the document set must be implemented. At this stage, the clustering provides further evidence of the inadequacy of the term independence model.

CHAPTER 6

CONCLUSIONS AND FURTHER RESEARCH

There have been many attempts to model bibliographic retrieval systems, which are complex systems with many components, each with many aspects to consider. This work has explored models useful for the simulation of these bibliographic retrieval systems and made some checks on their validity. The potential value of this approach is the greater understanding of how such systems work. By using simulation, the input parameters can be varied to check their effect on the performance of the system, but first a valid simulation model must be designed. This research is another step towards building a valid simulation model.

As with many modelling situations, there is a certain tradeoff between describing the system by a small number of parameters and capturing all the important elements of the system accurately. Other models such as the Bahadur-Lazarfield expansion (Yu, Luk, and Siu[128]), the Chow representation (Van Rijsbergen[119]), and the Linked-2-Poisson model (Bookstein and Kraft[10a]) have a

very large number of parameters to estimate for every database they are applied to, even though they incorporate term dependencies. The model presented here uses a much smaller number of parameters by using distributions of various components and by deriving the parameters of the co-occurrence distributions from the term occurrence distributions. The essential parts of a bibliographic retrieval system used for the simulation model in this study are the documents and their term assignments, queries and their term assignments, and the user judgements of documents. These components and their inter-relationships comprise the model explained in Chapter 2.

For the simulation model, first of all, the probability distributions of the objects in the system are needed. The split distribution in which a different type of distribution is used for the frequently occurring terms and the infrequent terms gives a better fit than a single distribution approach. This suggests that there is a different underlying phenomenon for the occurrence of the frequently occurring terms than for the less frequent terms. More work needs to be done in the area of finding appropriate goodness-of-fit tests for testing the distributions. This is a purely statistical question.

The simplest model of term assignment, i.e., independence of the assignment of index terms, does not fit,

as is shown by the co-occurrence distributions and the clustering experiments.

When term co-occurrence information was incorporated in a second-order probabilistic model, there was an improvement in the co-occurrence distribution of terms, but the recall-precision curves for the Medlars and Cranfield databases showed a very large discrepancy between the simulated data and the actual data. This is an indication that the dependence model does not go far enough in using term dependence information. Not only must second order dependence be incorporated but higher order dependence as well.

Although replicative validity has been tested where possible, more tests on large current databases such as CIJE need to be done. This is especially true for the query and user evaluation parts. Unfortunately, large complete information retrieval test collections including queries and evaluations as well as documents are few in number.

Only when such test collections are available can predictive validity be tested with some degree of success. Eventually, structural validity will also be tested, when a much greater understanding of the underlying mechanisms of bibliographic retrieval systems, particularly indexing and evaluation, is gained.

BIBLIOGRAPHY

1 Asai, Isao. "A General Formulation of Bradford's Distribution: The Graph Oriented Approach." Journal of the American Society for Information Science 32 (March 1981):113-119

2 Avramescu, Aurel. "Probabilistic Criteria for the Objective Design of Descriptor Languages." Journal of the American Society for Information Science 21 (March-April 1971):85-95

3 Baker, Frank B. "Latent Class Analysis as an Association Model for Information Retrieval." in Statistical Association Methods for Mechanized Documentation, pp.149-155. Edited by M.E. Stevens, V.E. Guiliano and L. Heilprin. National Bureau of Standards, 1965.

4 Baker, Norman R. and Nance, Richard E. "The Use of Simulation in Studying Information Storage and Retrieval Systems." American Documentation 19 (October 1968):363-370

5 Baxendale, P.B. "Machine Made Index for Technical Literature- An Experiment." IBM Journal of Research and Development 2 (1958):354-361

6 Bennett, John M. "Storage Design for Information Retrieval: Scarrott's Conjecture and Zipf's Law." International Computing Symposium 1975, pp.233-237. Amsterdam: North Holland, 1975.

7 Bird, P. R. "The Distribution of Indexing Depth in Documentation Systems." Journal of Documentation 30 (December 1974):381-390

8 Bird, P.R. "Some Sampling Characteristics of Bibliometric Distributions." Journal of Informatics 1 (August 1977):69-80

9 Bookstein, Abraham. "Statistical Behaviour of Search Keys." Journal of Library Automation 6 1973):109-116

10 Bookstein, Abraham and Cooper, William. "A General Mathematical Model for Information Retrieval Systems." Library Quarterly 46 (April 1976):153-167

10a Bookstein, Abraham and Kraft, Don. "Operations Research Applied to Document Indexing and Retrieval Decisions." Journal of the ACM 24 (July 1977):418-427

11 Bookstein, Abraham and Swanson, Don R. "Probabilistic Models for Automatic Indexing." Journal of the American Society for Information Science 25 (September-October 1974):312-318

12 Booth, Andrew D. "A 'Law' of Occurrences for Words of Low Frequency." Information and Control 10 (1967):386-393

13 Borko, Harold. "Studies on the Reliability and Validity for Factor-Analytically Derived Classification Categories." in Statistical Association Methods for Mechanized Documentation, pp.245-257. Edited by M.E. Stevens; V.E. Guiliano and L. Heilprin. National Bureau of Standards, 1965.

14 Bourne, C.P. and Ford, D.F. "Cost Analysis and Simulation Procedures for the Evaluation of Large Retrieval Systems." American Documentation 15 (April 1964):142-149

15 Brookes, B. C. "The Shannon Model of IR Systems." Journal of Documentation 28 (June 1972):160-162

16 Brookes, Bertram C. and Griffiths, Jose M. "Frequency Rank Distributions." Journal of the American Society for Information Science 29 (January 1978):5-13

17 Cagan , C. "A Highly Associative Document Retrieval System." Journal of the American Society for Information Science 21 (September-October 1970):330-337

18 Caras, Gus J. "Computer Simulation of a Small Information System." American Documentation 19 (April 1968):120-122

19 Chahil, Gurcharan Singh. "Simulation of Bibliographic Data Bases for Studies of Automatic Document Classification." M.Sc. Dissertation, Concordia University, Montreal, 1979.

20 Chapman, Robert L. "The Case for Information Systems Simulation." in Information System Sciences, pp.477-484. Edited by J. Spiegel and D. Walker. Washington, D.C.: Spartan Books, 1965.

21 Chow, K.K. and Liu, C.N. "Approximating Discrete Probability Distributions with Dependence Trees." IEEE Transactions on Information Theory IT-14 (May 1968):462-467

22 Cigler, Ing K. "Simulation Methods Applied in Designing an Information System," Information Storage and Retrieval 6 (October 1970):307-312

23 Cleveland, Donald B. "An N-Dimensional Retrieval Model." Journal of the American Society for Information Science 27 (September-October 1976):342-347

24 Cleverdon, C.W. and Keen, E.M. Factors Determining the Performance of Indexing Systems. 2 vols. Cranfield, England: Aslib Cranfield Research Project, 1966.

25 Comer, Douglas. "The Difficulty of Optimum Index Selection." ACM Transaction on Database Systems 3 (December 1978):440-445

26 Conover, W.J. Practical Nonparametric Statistics. 2 ed. New York: John Wiley, 1980.

27 Cooper, Michael D. "A Simulation Model of an Information Retrieval System." Information Storage and Retrieval 9 (January 1973):13-32

28 Croft, W.B. and Harper, D.J. "Using Probabilistic Models of Document Retrieval without Relevance Information." Journal of Documentation 35 (December 1979):285-295

29 Curtice, Robert M. and Jones, Paul E. "Distributional Constraints and Automatic Selection of an Indexing Vocabulary." Proceedings of the American Documentation Institute Annual Meeting. 4 (1967):152-156

30 Dabrowski, Miroslaw. "A General Model of Distribution of Objects in Information Retrieval Systems." Information Systems 1 (1975):147-151

31 De Lutis, Thomas G.; Rush, J.E. and Wong, P. "The Modelling of a Large On-Line Real Time Information System." Simulation Symposium 1977, pp.353-370. Tampa, Florida: Annual Simulation Symposium, 1977.

32 Doszkocs, Tamas E. "AID, an Associative Interactive Dictionary for Online Searching." Online Review 2 (June 1978):163-173

33 Doszkocs,Tamas ; Schultheisz, Robert J. and Vasta, Bruno, M. "Analysis of Term Distribution in the Toxline Inverted File." Journal of Chemical Information and Computer Sciences 16 (August 1976):131-135

34 Doyle, L.B. "Association Characteristics of Words in Text." Communications of the ACM 5 (April 1962):223

35 Duda, Richard O. and Hart, Peter E. Pattern Classification and Scene Analysis. New York: Wiley, 1973.

36 Fairthorne, R.A. "Algebraic Representation of Storage and Retrieval Languages." in Proceedings of the International Conference on Scientific Information, pp.1313-1326. Washington, D.C.: National Academy of Sciences, 1959.

37 Fairthorne, Robert A. "Empirical Hyperbolic Distributions (Bradford Zipf Mandelbrot) for Bibliometric Description and Prediction." Journal of Documentation 25 (December 1969):319-343

38 Fishman, George S. Principles of Discrete Event Simulation. New York:Wiley, 1978.

39 Fried, J.B. et al. Index Simulation Feasibility and Automatic Document Classification. Columbus, Ohio: Ohio State University, 1968.

40 Gebhardt, Friedrich. "A Simple Probabilistic Model for the Relevance Assessment of Documents." Information Processing and Management 11 (June 1975):59-65

41 Goffman, William. "An Indirect Method of Information Retrieval." Information Storage and Retrieval 4 (1969):361-373

42 Good, I.J. "Statistics of Language : Introduction." Encyclopedia of Linguistics, Information and Control, pp.567-581. Edited by R.A. Meetham. Oxford: Pergamon Press, [1969].

43 Griffiths, Jose-Marie. "Computer Simulation: A Research Tool for Information Science." in New Trends in Documentation and Information, Proceedings of the 39th FID Congress 1978, pp.137-144. London: ASLIB, 1980.

44 Griffiths, Jose-Marie. "Index Term Input to IR Systems." Journal of Documentation 31 (September 1975):185-190

45 Griffiths, Jose-Marie. "The Computer Simulation of Information Retrieval Systems." Ph.D. dissertation. London: University College, 1977.

46 Harper, D.J. "Using Term Dependence Information in Document Retrieval." Journal of Informatics 2 (April 1978):82

47 Harper, D.J. and Van Rijsbergen, C.J. "An Evaluation of Feedback in Document Retrieval Using Co-occurence Data." Journal of Documentation 34 (September 1978):189-216

48 Harter, Stephen P. "A Probabilistic Approach to Automatic Keyword Indexing, Part I On the Distribution of Specialty Words in a Technical Literature." Journal of the American Society for Information Science 26 (July-August 1975):197-206

49 Harter, Stephen P. "A Probabilistic Approach to Automatic Keyword Indexing Part II An Algorithm for Probabilistic Indexing." Journal of the American Society for Information Science 26 (September-October 1975):280-289

50 Hill, Bruce M. "The Rank-Frequency Form of Zipf's Law." Journal of the American Statistical Association 69 (December 1974):1017-1026

51 Hill, Bruce M. and Woodroofe, Michael. "Stronger Forms of Zipf's Law." Journal of the American Statistical Association 70 (January 1975):212-219

52 Hillman, D.J. "Mathematical Classification Techniques for Non-Static Collections." in Classification Research, pp.177-209. Edited by P. Atherton. Copenhagen: Munksgaard, 1965.

53 Hodes, Louis. "Selection of Descriptors According to Discrimination and Redundancy . Application to Chemical Structure Searching." Journal of Chemical Information and Computer Sciences 16 (May 1976):88-93

54 Horn, Susan Dadkis. "Goodness of Fit Tests for Discrete Data: A Review and an Application to a Health Impairment Scale." Biometrics 33 (March 1977):237-248

55 Houston, N. and Wall, E. "The Distribution of Term Usage in Manipulative Indexes." American Documentation 15 (April 1964):105-114

56 Hubert, J.J. "Analysis of Data by a Rank Frequency Model." Ph.D. dissertation, SUNY at Buffalo, 1974.

57 Hubert, John J. "A Relationship Between Two Forms of Bradford's Law." Journal of the American Society for Information Science 29 (May 1978):159-161

58 Jacquesson, Alain and Schieber, William D. "Term Association Analysis on a Large File of Bibliographic Data, Using a Highly Controlled Indexing Vocabulary." Information Storage and Retrieval 9 (February 1973):85-94

59 Johnson, N.L. and Kotz, S. Distributions in Statistics: Discrete Distributions. Boston: Houghton Mifflin, 1969.

60 Krevitt, Beth and Grifith, Belver C. "A Comparison of Several Zipf Type Distributions in their Goodness of Fit to Language Data." Journal of the American Society for Information Science 23 (May-June 1972):220-221

61 Lancaster, W.F. "Interaction Between Requestors and a Large Mechanized Retrieval System." Information Storage and Retrieval 4 (1968):239-252

62 Laus, Krystyna and Dabrowski, Myroslaw. "A Model of the Information Retrieval Process for Hierarchial Sets of Descriptors." Information Storage and Retrieval 10 (July-August 1974):261-265

63 Lesk, M.E. "Word-word Associations in Document Retrieval Systems." American Documentation 20 (January 1969):27-38

64 Liebsney, Felix. State of the Art Survey on Automatic Indexing. Paris: UNESCO, 1974. COM/WS/18

65 Lowe, T.C. "The Influence of Data Base Characteristics and Usage on Direct File Organization." Journal of the ACM 15 (October 1968):534-548

66 Ludwig, B.M. and Glockman, H.P. "The Formal Analysis of Document Retrieval Systems." Journal of the American Society for Information Science 26 (January-February 1975):51-55

67 Luhn, H.P. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information." IBM Journal Research and Development 1 (1957):309-317

68 MacMillan, R.D. Chris. "The Simulation of an Automated Information Retrieval System." M.Sc. Dissertation. Kingston, Ontario: Queens University, 1976.

69 Mandelbrot, Benoit. "An Informational Theory of the Statistical Structure of Language." in Communication Theory, pp.486-502. Edited by Willis Jackson. London: Butterworths Scientific Publications, 1953.

70 Marek, Wiktor and Pawlak, Zdzislaw. Information Storage and Retrieval System - Mathematical Foundations. CC Pas. Reports 149, Warsaw, 1974.

71 Maron, M.E. "Depth of Indexing." Journal of the American Society for Information Science 30 (July 1979):224-228

72 Maron, M.E. and Kuhns, J.L. "On Relevance, Probabilistic Indexing and Information Retrieval." Journal of the ACM 7 (July 1960):295-311

73 Mazur, Zygmunt. "Inverted File Organization in the Information Retrieval System Based on Thesaurus with Weights." Information Processing and Management 15 (1979):227-234

74 Mazur, Zygmunt. "Properties of a Model of Information Retrieval System Based on Thesarurus with Weights." Information Processing and Management 15 (1979):145-154

75 McBirney, Warren B. Frequency of Term use in Indexing with a Controlled Vocabulary. Washington, D.C.: U.S. Dept. of the Interior, 1970. PB195742

76 Miller, William L. "A Probabilistic Search Strategy for Medlars." Journal of Documentation 27 (December 1974):254-266

77 Mooers, C.N. "A Mathematical Theory of the Use of Language Symbols in Retrieval." in Proceedings of the International Conference on Scientific Information, pp.1327-1364. Washington, D.C.: National Academy of Sciences, 1959.

78 Nelson, M.J. "Distributions of Keywords and their Co-occurrences." Unpublished Report, School of Library and Information Science, University of Western Ontario, 1978.

79 O'Neill Edward T. and Llinas, James. "A Method for Evaluating Search Key Performance." Proceedings of the American Society for Information Science Annual Meeting. 13 (1976):207-221

80 Radecki, Tadeusz. "Fuzzy Set Theoretical Approach to Document Retrieval." Information Processing and Management 15 (1979):247-260

81 Radecki, Tadeusz. "Mathematical Model of Information Retrieval System Based on the Concept of Fuzzy Thesaurus." Information Processing and Management 12 (1976):313-318

82 Raghavan, Vijay V. and Yu, C.T. "Experiments on the Determination of the Relationships between Terms." ACM Transactions on Database Systems 4 (June 1979):240-260

83 Rastogi, Kunj B. "Retrieval Behavior of Derived Truncated Search Keys for a Large On-Line Bibliographic File." Journal of the American Society for Information Science 31 (March 1980):84-88

84 Raver, Norman. "Performance of IR Systems." in Information Retrieval: A Critical View, pp.131-142. Edited by George Schecter. Washington, D.C.; Thompson Book, 1967.

85 Roberts, Fred S. Discrete Mathematical Models. Englewood Cliffs, New Jersey: Prentice-Hall, 1976.

86 Robertson, S.E. and Sparck Jones, K. "Relevance Weighting of Search Terms." Journal of the American Society for Information Science 27 (May-June 1976):129-146

87 Robertson, Stephen E. "Theories and Models in Information Retrieval." Journal of Documentation 33 (June 1977):126-148

88 Salton, G. "Computer Evaluation of Indexing and Text Processing". Journal of the ACM 15 (1968):8-36

89 Salton, G. A Theory of Indexing. Philadelphia: Society for Industrial and Applied Mathematics, 1975.

90 Salton, G. Automatic Information Organization and Retrieval. New York: McGraw-Hill, 1968.

91 Salton, G. "Mathematics and Information Retrieval." Journal of Documentation 35 (March 1979):1-29

92 Salton, G.; Wong, A. and Yang, C.S. "A Vector Space Model for Automatic Indexing." Communications of the ACM 18(11) (November 1975):613-620

93 Salton, G.; Wong, A. and Yu, C.T. "Automatic Indexing Using Term Discrimination and Term Precision Measurements." Information Processing and Management 12 (1976):43-51

94 Saracevic, T. An Inquiry into Testing of Information Retrieval Systems. 3 vols. Cleveland: Center for Documentation and Communications Research, Case Western Reserve University, 1968.

95 Schuegraf, E.J. "Compression of Large Inverted Files with Hyperbolic Term Distribution." Information Processing and Management 12 (1976):377-384

96 Schultz, Claire K.; Schwartz, Phyllis D. and Steinberg, Leon. "A Comparison of Dictionary Use Within Two Information Retrieval Systems." American Documentation 12 (October 1961):247-253

97 Sichel, H.S. "On a Distribution Law for Word Frequencies." Journal of the American Statistical Association 70 (1975):542-548

98 Siler, Kenneth F. "A Stochastic Evaluation Model for Database Organizations in Data Retrieval Systems." Communications of the ACM 19 (February 1976):84-95

99 Simon, H.A. "On a Class of Skew Distribution Functions." Biometrika 42 (1955):425-440

100 Soergel, Dagobert. "Mathematical Analysis of Documentation Systems." Information Storage and Retrieval 3 (July 1967):129-173

101 Sparck Jones, Karen. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." Journal of Documentation 28 (March 1972):11-21

102 Sparck Jones, Karen. "Experiments in Relevance Weighting of Search Terms." Information Processing and Management 15 (1979):133-144

103 Stevens, Mary Elizabeth; Giuliano, Vincent E. and Heilprin, Laurence (eds.). Statistical Association Methods for Mechanized Documentation. National Bureau of Standards miscellaneous publication 269, 1965.

104 Stiles, H.E. "The Association Factor in Information Retrieval." Journal of the ACM 8 (April 1961):271-279

105 Svenonius, Elaine. "An Experiment in Index Term Frequency." Journal of the American Society for Information Science 23 (March-April 1972):109-121

106 Swanson, Don R. "On Indexing Depth and Retrieval Effectiveness." in Information System Sciences, pp.311-319. Edited by J. Spiegel and D. Walker. Washington, D.C.: Spartan Books, 1965.

107 Switzer, Paul. "Vector Images in Document Retrieval." in Statistical Association Methods for Mechanized Documentation, pp.163-171. Edited by M.E. Stevens; V.E. Guiliano and L. Heilprin. National Bureau of Standards, 1965.

108 Tague, J. "Simulation of Information Retrieval Systems." Unpublished Report, School of Library and Information Science, University of Western Ontario, 1977.

109 Tague, J.M. "A Bayesian Approach to Interactive Retrieval." Information Storage and Retrieval 9 (March 1973):129-142

110 Tague, J.M. "User-Responsive Subject Control in Bibliographic Retrieval Systems". Information Processing and Management 17 (1981):149-159

111 Tague, J.M. and Nelson, M.J. "Simulation of User Judgments in Bibliographic Retrieval Systems." In Proceedings of the Fourth International Conference on Information Retrieval. FORUM 16 (Summer 1981):66-71

112 Tague, J.;Nelson, M.J. and Wu, H. "Problems in the Simulation of Bibliographic Retrieval Systems." In Information Retrieval Research. Edited by R.N. Oddy et al. London: Butterworths, 1981.

113 Tahani, Valiollah. "A Fuzzy Set Model of Document Retrieval Systems." Information Processing and Management 12 (1976):177-187

114 Takahama, Tadahiko. "A Model for a Document Retrieval System." Information Storage and Retrieval 9 1973:143-163

115 Toma, Eugeniu. "The Structure of the Euratom Thesaurus." Journal of Documentation 27 (December 1974):267-272

116 Turski, W.M. "On a Model of Information Retrieval System Based on Thesaurus." Information Storage and Retrieval 7 (August 1971):89-94

117 Tyron, R.C. and Bailey, D.E. Cluster Analysis. New York: McGraw-Hill, 1970.

118 Uhlmann, Wolfram. "Document Specification and Search Strategy Using Basic Intersections and the Probability Measure of Sets." American Documentation 19 (July 1968):240-246

119 Van Rijsbergen, C.J. "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval." Journal of Documentation 33 (June 1977):106-119

120 Van Rijsbergen, C.J.; Harper, D.J. and Porter, M.F. "The Selection of Good Search Terms." Information Processing and Management 17 (1981):77-91

121 Wall, Eugene. "Further Implications of the Distribution of Index Term Usage." Proceedings of the American Documentation Institute Annual Meeting. 1 (1964):457-466

122 Waller, W.G. and Kraft, Donald H. "A Mathematical Model of a Weighted Boolean Retrieval System." Information Processing and Management 15 (1979):235-246

123 Walsh, John E. "The Poisson Distribution as a Limit for Dependent Binomial Events with Unequal Probabilities." Operations Research 3 (1955):198-204

124 Willett, Peter. "A Fast Procedure for the Classification of Similarity Coefficients in Automatic Classification." Information Processing and Management 17 (1981):53-68

125 Williams, J.H. "Results of Classifying Documents with Multiple Discriminant Functions." in Statistical Association Methods for Mechanized Documentation, pp. 217-224. Edited by M.E. Stevens; V.E. Guiliano and L. Heilprin. National Bureau of Standards, 1965.

126 Williams, Martha E. ed. Computer-Readable Data Bases: A Directory and Data Source Book. Washington, D.C.: American Society for Information Science, 1979.

127 Wishart, D. CLUSTAN User Manual: 3 ed. Edinburgh: Program Library Unit, Edinburgh University, 1978.

128 Yu, C.T.; Luk, W.S. and Siu, M.K. "On Models of Information Retrieval Processes." Information Systems 4 (1979):205-218

129 Yu, C.T.; Luk, W.S. and Siu, M.K. "On the Estimation of the Number of Desired Records with Respect to a Given Query." ACM Transactions on Database Systems 33 (March 1978):41-56

130 Yu, C.T. and Salton, G. "Precision Weighting- An Effective Automatic Indexing Method." Journal of the ACM 23 (January 1976):77-88

131 Zeigler, B.P. Theory of Modelling and Simulation. New York: Wiley, 1976.

132 Zgrzywa, Aleksander. "Application of Computer Simulation Methods to the Investigation of On-line Information Retrieval Systems." in New Trends in Documentation and Information, Proceedings of the 39th FID Congress 1978, pp.145-157. London: ASLIB, 1980.

133 Zipf, G.K. Human Behaviour and the Principle of Least Effort. Cambridge, Mass.:Addison-Wesley, 1949.

134 Zunde, Pranas. "Predictive Models of Information Systems." Information Processing and Management 17 (1981):103-111

135 Zunde, Pranas and Slamecka, Vladimir. "Distribution of Indexing Terms for Maximum Efficiency of Information Transmission." American Documentation 18 (April 1967):104-108

APPENDIX I

THE SIMULATION PROGRAMS

The simulation programs were written in FORTRAN (Control Data FTN4) for the CYBER-73 at the University of Western Ontario Computing Center. The original versions were programmed by Harry Wu at Cornell University for an IBM-370. They were converted to run on the CYBER-73 and many modifications and additions were made.

The programs have as input several parameters which describe a bibliographic retrieval system (see Chapter 4). Using these parameters the programs simulate the retrieval system document by document. First the number of index terms which are assigned to a document is calculated by randomly sampling from the distribution of exhaustivity. Then the first index term is randomly sampled from the split distribution of index terms. This is accomplished by converting the frequency-size part of the distribution to a frequency-rank and using the total distribution of indexing terms as a frequency-rank. The rank of a term is its unique identifier as terms with equal frequency are given unique ranks, assigned arbitrarily.

For the independent model each term is randomly selected from this frequecy-rank distribution. For the dependent model after the initial term is selected, the

conditional probabilities of all other terms occurring given the frequency of the initial term, are calculated and stored. This conditional probability distribution is used to select all subsequent terms for the document until the correct number is reached. This final document vector is then written to a disk file for further processing and statistics.

The program has a main routine called SIMULAT, two functions called XICOMP and GENNUM, and three subroutines GENTAB, GENVEC, GENCO. The function GGUBFS from the IMSL subroutine package was used as a random number generator.

SIMULAT has the following main components:

1. calculates the frequecy-rank part of the term distribution,
2. calculates the frequency-size part of the term distribution and converts the term probabilites to a frequency-rank form for combining with the frequency rank portion,
3. calculates the joint probabilities and stores them in a random access disk file,
4. uses GENTAB to generate a table of the distribution function of exhaustivity probabilities for documents and queries,
5. repeatedly calls GENNUM to sample the exhaustivity for generating a document length DL (number of terms assigned),

and calls GENVEC to generate a document vector of length DL, until the required number of documents are generated.

XICOMP: calculates the parameter "k" for the Zipf distribution.

GENTAB: given the distribution parameters, generates a distribution function in a table.

GENNUM: given a random number between zero and one, it does a binary search of a distribution table to find a randomly sampled value.

GENVEC: given a vector length DL, it generates DL terms to be assigned to a document. In version 1 this is accomplished by successive samples from a distribution function using GENNUM. In version 2, the dependent model, it first calls GENCO to generate the first term, then generates the second and subsequent terms by using the conditional probabilites.

GENCO: first generates the first terms randomly using GENNUM and generates the conditional probabilities of all other terms given this first term by using the joint probabilites stored on disk in SIMULAT.

More detailed algorithms for many parts of the program can be found in Tague, Nelson, and Wu[112].

TABLE 2

TERM OCCURRENCES - Medlars

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
1	2598	1	138	138
2	640	2	132	270
3	362	3	120	390
4	245	4	99	489
5	151	5	95	584
6	90	8	82	830
7	85	9	76	906
8	70	10	74	980
9	62	12	69	1118
10	51	13	68	1186
11	45	14	61	1247
12	31	16	60	1367
13	21	17	58	1425
14	28	18	55	1480
15	19	20	53	1586
16	22	21	52	1638
17	14	22	51	1689
18	17	23	50	1739
19	12	24	49	1788
20	9	25	48	1836
21	11	26	47	1883
22	10	30	46	2067
23	9	34	45	2247
24	7	36	44	2335

TABLE 2 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
25	7	37	43	2378
26	9	39	41	2460
27	6	41	40	2540
28	3	44	39	2657
29	4	53	37	2990
30	8	55	36	3062
31	5	60	35	3237
32	8	65	34	3407
33	2	67	33	3473
34	5	75	32	3729
35	5	80	31	3884
36	2	88	30	4124
37	9	92	29	4240
39	3	95	28	4324
40	2	101	27	4486
41	2	110	26	4720
43	1	117	25	4895
44	2	124	24	5063
45	4	133	23	5270
46	4	143	22	5490
47	1	154	21	5721
48	1	163	20	5901
49	1	175	19	6129
50	1	192	18	6435
51	1	206	17	6673

TABLE 2 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
52	1	228	16	7025
53	2	247	15	7310
55	1	275	14	7702
58	1	296	13	7975
60	2	327	12	8347
61	1	372	11	8842
68	1	423	10	9352
69	2	485	9	9910
74	1	555	8	10470
76	1	640	7	11065
82	3	730	6	11605
95	1	881	5	12360
99	1	1126	4	13340
120	1	1488	3	14426
132	1	2128	2	15706
138	1	4726	1	18304

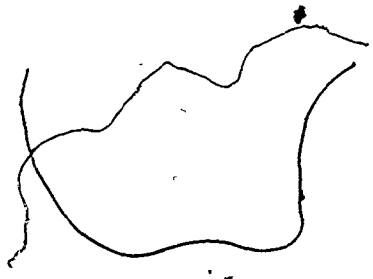


TABLE 3

TERM OCCURRENCES - CRANFIELD

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
1	1132	2	214	428
2	350	3	192	620
3	186	4	160	780
4	140	5	159	939
5	91	6	155	1094
6	81	7	149	1243
7	70	8	148	1391
8	51	9	137	1528
9	52	11	134	1796
10	34	12	132	1928
11	34	13	129	2057
12	21	15	128	2313
13	14	16	119	2432
14	18	17	115	2547
15	22	18	112	2659
16	17	19	106	2765
17	9	21	105	2975
18	19	23	104	3183
19	10	24	103	3286
20	15	25	101	3387
21	12	26	100	3487
22	16	27	99	3586
23	11	29	98	3782
24	2	31	97	3976

TABLE 3 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
25	11	33	95	4166
26	11	34	92	4258
27	12	37	91	4531
28	6	38	90	4621
29	8	40	88	4797
30	3	41	87	4884
31	13	43	85	5054
32	6	44	83	5137
33	4	45	82	5219
34	3	46	79	5298
35	12	48	76	5450
36	2	52	75	5750
37	5	53	74	5824
38	6	54	73	5897
39	6	55	72	5969
40	8	56	71	6040
41	2	60	70	6320
42	9	61	69	6389
44	1	62	68	6457
45	4	63	67	6524
46	2	65	66	6656
47	3	67	65	6786
48	3	68	63	6849
49	4	69	62	6911
50	3	71	60	7031

TABLE 3' (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
52	1	75	59	7267
53	3	78	58	7441
54	3	83	57	7726
55	2	88	56	8006
56	5	90	55	8116
57	5	93	54	8278
58	3	96	53	8437
59	4	97	52	8489
60	2	100	50	8639
62	1	104	49	8835
63	1	107	48	8979
65	2	110	47	9120
66	2	112	46	9212
67	1	116	45	9392
68	1	117	44	9436
69	1	126	42	9814
70	4	128	41	9896
71	1	136	40	10216
72	1	142	39	10450
73	1	148	38	10678
74	1	153	37	10863
75	4	155	36	10935
76	2	167	35	11355
79	1	170	34	11457
82	1	174	33	11589

TABLE 3 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
83	1	180	32	11781
85	2	193	31	12184
87	1	196	30	12274
88	2	204	29	12506
90	1	210	28	12674
91	3	222	27	12998
92	1	233	26	13284
95	2	244	25	13559
97	2	246	24	13607
98	2	257	23	13860
99	1	273	22	14212
100	1	285	21	14464
101	1	300	20	14764
103	1	310	19	14954
104	2	329	18	15296
105	2	338	17	15449
106	1	355	16	15721
112	1	377	15	16051
115	1	395	14	16303
119	1	409	13	16485
128	2	430	12	16737
129	1	464	11	17111
132	1	498	10	17451
134	2	550	9	17919
137	1	601	8	18327

TABLE 3 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
148	1	671	7	18817
149	1	752	6	19303
155	1	843	5	19758
159	1	983	4	20318
160	1	1169	3	20876
192	1	1519	2	21576
214	2	2651	1	22708

TABLE 4.

TERM OCCURRENCES - SPI

FREQUENCY-SIZE x	g(x)	FREQUENCY-RANK r	f(r)	F(r)
1	723	1	127	127
2	154	2	87	214
3	74	3	53	267
4	57	4	36	303
5	31	5	34	337
6	22	6	26	363
7	12	8	23	409
8	7	9	21	430
9	14	12	20	490
10	7	15	19	547
11	12	16	18	565
12	6	21	17	650
13	3	24	16	698
14	3	25	15	713
15	1	28	14	755
16	3	31	13	794
17	5	37	12	866
18	1	49	11	998
19	3	56	10	1068
20	3	70	9	1194
21	1	77	8	1250
23	2	89	7	1334
26	1	111	6	1466
34	1	142	5	1621

TABLE 4 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
36	1	199	4	1849
53	1	273	3	2071
87	1	427	2	2379
127	1	1150	1	3102

TABLE 5

TERM OCCURRENCES - CIJE

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
1	482	1	2406	2406
2	341	2	1567	3973
3	287	3	1069	5042
4	222	4	1037	6079
5	193	5	998	7077
6	196	6	610	7687
7	169	7	586	8273
8	135	8	549	8822
9	119	9	475	9297
10	127	10	420	9717
11	96	11	419	10136
12	105	12	387	10523
13	82	13	381	10904
14	76	14	363	11267
15	77	15	355	11622
16	64	16	353	11975
17	51	17	352	12327
18	48	18	330	12657
19	59	19	327	12984
20	41	20	318	13302
21	30	21	313	13615
22	44	22	302	13917
23	43	23	292	14209
24	31	24	287	14496

TABLE 5 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
25	42	25	284	14780
26	39	26	279	15059
27	42	27	270	15329
28	25	28	266	15595
29	22	29	262	15857
30	33	30	259	16116
31	27	32	251	16618
32	29	33	246	16864
33	32	34	240	17104
34	27	35	231	17335
35	16	36	230	17565
36	18	38	225	18015
37	21	39	218	18233
38	16	40	213	18446
39	12	41	209	18655
40	24	42	205	18860
41	17	43	202	19062
42	19	44	199	19261
43	14	45	195	19456
44	16	46	193	19649
45	11	48	191	20031
46	14	50	188	20407
47	13	52	187	20781
48	16	54	186	21153
49	10	56	185	21523

TABLE 5 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
50	13	58	183	21889
51	10	59	182	22071
52	9	60	181	22252
53	9	61	180	22432
54	14	62	178	22610
55	9	63	176	22786
56	9	64	175	22961
57	8	66	173	23307
58	4	68	169	23645
59	10	69	167	23812
60	5	72	166	24310
61	9	76	164	24966
62	7	77	162	25128
63	7	78	161	25289
64	5	80	158	25605
65	11	81	157	25762
66	5	82	153	25915
67	10	83	152	26067
68	10	85	148	26363
69	7	86	146	26509
70	3	87	145	26654
71	3	88	143	26797
72	8	89	142	26939
73	10	91	141	27221
74	4	92	140	27361

TABLE 5 (Continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
75	4	93	139	27500
76	6	94	138	27638
77	4	95	136	27774
78	2	98	135	28179
79	5	102	133	28711
80	5	105	132	29107
81	5	108	131	29500
82	2	111	130	29890
83	3	112	128	30018
84	4	114	127	30272
85	4	115	125	30397
86	3	118	124	30769
87	8	122	123	31261
88	3	124	122	31505
89	7	130	121	32231
90	5	131	120	32351
91	5	132	119	32470
92	4	134	118	32706
93	6	138	117	33174
94	8	141	114	33516
95	3	142	113	33629
96	1	145	112	33965
97	3	147	111	34187
98	1	149	110	34407
99	5	155	109	35061

TABLE 5 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
100	5	157	108	35277
101	2	159	107	35491
102	1	160	106	35597
103	3	166	105	36227
104	2	168	104	36435
105	6	171	103	36744
106	1	172	102	36846
107	2	174	101	37048
108	2	179	100	37548
109	6	184	99	38043
110	2	185	98	38141
111	2	188	97	38432
112	3	189	96	38528
113	1	192	95	38813
114	3	200	94	39565
117	4	206	93	40123
118	2	210	92	40491
119	1	215	91	40946
120	1	220	90	41396
121	6	227	89	42019
122	2	230	88	42283
123	4	238	87	42979
124	3	241	86	43237
125	1	245	85	43577
127	2	249	84	43913

TABLE 5 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
128	1	252	83	44162
130	3	254	82	44326
131	3	259	81	44731
132	3	264	80	45131
133	4	269	79	45526
135	3	271	78	45682
136	1	275	77	45990
138	1	281	76	46446
139	1	285	75	46746
140	1	289	74	47042
141	2	299	73	47772
142	1	307	72	48348
143	1	310	71	48561
145	1	313	70	48771
146	1	320	69	49254
148	2	330	68	49934
152	1	340	67	50604
153	1	345	66	50934
157	1	356	65	51649
158	2	361	64	51969
161	1	368	63	52410
162	1	375	62	52844
164	4	384	61	53393
166	3	389	60	53693
167	1	399	59	5[illegible]83

TABLE 5 (continued)

FREQUENCY-SIZE x	g(x)	FREQUENCY-RANK r	f(r)	F(r)
169	2	403	58	54515
173	2	411	57	54971
175	1	420	56	55475
176	1	429	55	55970
178	1	443	54	56726
180	1	452	53	57203
181	1	461	52	57671
182	1	471	51	58181
183	2	484	50	58831
185	2	494	49	59321
186	2	510	48	60089
187	2	523	47	60700
188	2	537	46	61344
191	2	548	45	61839
193	1	564	44	62543
195	1	578	43	63145
199	1	597	42	63943
202	1	614	41	64640
205	1	638	40	65600
209	1	650	39	66068
213	1	666	38	66676
218	1	687	37	67453
225	2	705	36	68101
230	1	721	35	68661
231	1	748	34	69579

TABLE 5 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
240	1	780	33	70635
246	1	809	32	71563
251	2	836	31	72400
259	1	869	30	73390
262	1	891	29	74028
266	1	916	28	74728
270	1	958	27	75862
279	1	997	26	76876
284	1	1039	25	77926
287	1	1070	24	78670
292	1	1113	23	79659
302	1	1157	22	80627
313	1	1187	21	81257
318	1	1228	20	82077
327	1	1287	19	83198
330	1	1335	18	84062
352	1	1386	17	84929
353	1	1450	16	85953
355	1	1527	15	87108
363	1	1603	14	88172
381	1	1685	13	89238
387	1	1790	12	90498
419	1	1886	11	91554
420	1	2013	10	92824
475	1	2132	9	93899

TABLE 5 (continued)

FREQUENCY-SIZE		FREQUENCY-RANK		
x	g(x)	r	f(r)	F(r)
549	1	2267	8	94975
586	1	2436	7	96158
610	1	2632	6	97334
998	1	2825	5	98299
1037	1	3047	4	99187
1069	1	3334	3	100048
1567	1	3675	2	100730
2406	1	4157	1	101212

TABLE 6

FIT OF TERM DISTRIBUTIONS - MEDLARS

A. Frequency-size segment - Mandelbrot-Zipf.

a=4448.8 b=2.0121 c=0.3396 H1=48

Number of Term Occurrences (x)	Number of Terms with x Occurrences Observed	Expected	Cumulative Chi-square
1	2598	2470.328	6.598
2	640	804.436	40.211
3	362	393.111	42.673
4	245	232.075	43.393
5	151	152.905	43.417
6	90	108.246	46.492
7	85	80.616	46.731
8	70	62.345	47.671
9	62	49.641	50.747
10 - 11	96	74.050	57.254
12 - 13	52	52.572	57.260
14 - 16	69	55.351	60.626
17 - 20	52	48.909	60.821
21 - 25	44	39.749	61.276
26 - 33	45	39.059	62.180
34 - 48	41	38.606	62.328

TABLE 6 (continued)

B. Frequency-rank segment - Log-rank

a=-28.1 b=138 H2=24

Rank of Term	Number of Term Occurrences	
	Observed	Expected
1	138	138
2	132	118.5
3	120	107.1
4	99	99.0
5	95	92.7
6	82	87.6
7	82	83.3
7	82	79.5
9	76	76.2
10	74	73.2
11	69	70.5
12	68	65.8
13	68	65.8
14	61	63.8
15	60	61.8
16	60	60.0
17	58	58.3
18	55	56.7
19	53	55.2
20	53	53.7
21	52	52.4
22	51	51.0
23	50	49.7
24	49	48.6

TABLE 7

FIT OF TERM DISTRIBUTIONS - CRANFIELD

A. Frequency-size segment - Mandelbrot-Zipf.

a=1072.7 b=1.488 c=0.0 H1=111

Number of Term Occurrences (x)	Number of Terms with x Occurrences		Cumulative Chi-square
	Observed	Expected	
1	1132	1072.744	3.273
2	350	382.440	6.025
3	186	209.189	8.595
4	140	136.342	8.694
5	91	97.820	9.169
6	81	74.577	9.722
7	70	59.291	11.656
8	51	48.607	11.774
9	52	40.793	14.853
10	34	34.874	14.875
11	34	30.262	15.337
12 - 13	35	50.189	19.934
14 - 15	40	40.213	19.935
16 - 17	26	33.163	21.482
18 - 20	44	40.394	21.804
21 - 23	39	32.449	23.126
24 - 27	36	34.767	23.170
28 - 32	36	34.143	23.271
33 - 38	32	31.899	23.271
39 - 45	30	28.975	23.307
46 - 54	22	28.762	24.897

TABLE 7 (continued)

55 - 66	27	28.910	25.024
67 - 83	20	29.806	28.250
84 - 111	25	33.388	30.357

B. Frequency-rank segment - Log-rank

a=-33.4 b=214 H2=18

Rank of Term	Number of Term Occurrences Observed	Expected
1	214	214
2	214	190.8
3	192	177.3
4	160	167.7
5	159	160.3
6	155	154.2
7	149	149.0
8	148	144.6
9	137	140.6
10	134	137.1
11	134	133.9
12	132	131.0
13	129	128.4
14	128	125.9
16	128	123.6
16	119	121.4
17	115	119.4
18	112	117.5

TABLE 8

FIT OF TERM DISTRIBUTIONS - SPI

A. Frequency-size segment - Mandelbrot-Zipf.

a=711.94 b=2.00 c=0.0 H1=25

Number of Term Occurrences (x)	Number of Terms with x Occurrences Observed	Expected	Cumulative Chi-square
1	723	711.943	0.172
2	154	178.393	3.507
3	74	79.392	3.873
4	57	44.700	7.258
5	31	28.629	7.454
6	22	19.893	7.677
7	12	14.623	8.148
8 - 9	21	20.054	8.192
10 - 12	25	18.089	10.833
13 - 16	10	13.912	11.933
17 - 25	15	15.371	11.941

B. Frequency-rank segment - Log-rank

a=-60.1 b=127 H2=6

Rank of Term	Number of Term Occurrences Observed	Expected
1	127	127
2	87	85.3
3	53	60.9
4	36	43.6
5	34	30.2
6	26	19.3

TABLE 9

FIT OF TERM DISTRIBUTIONS - CIJE

A. Frequency-size segment - Mandelbrot-Zipf.

a=12539.2 b=1.752 c=5.335 Hl=199

Number of Term Occurrences (x)	Number of Terms with x Occurrences		Cumulative Chi-square
	Observed	Expected	
1	482	493.867	0.285
2	341	382.023	4.690
3	287	305.382	5.797
4	222	250.397	9.017
5	193	209.507	10.318
6	196	178.206	12.094
7	169	153.672	13.623
8	135	134.055	13.630
9	119	118.103	13.637
10	127	104.943	18.278
11	96	93.948	18.318
12	105	84.659	23.205
13	82	76.736	23.566
14	76	69.919	24.095
15	77	64.007	26.732
16	64	58.844	27.184
17	51	54.306	27.385
18	48	50.294	27.490
19	59	46.730	30.712
20	41	43.546	30.861
21	30	40.691	33.670

TABLE 9 (continued)

22	44	38.119	34.577
23	43	35.793	36.028
24	31	33.683	36.242
25	42	31.762	39.542
26	39	30.007	42.237
27	42	28.400	48.750
28	25	26.925	48.887
29	22	25.566	49.384
30	33	24.312	52.489
31	27	23.152	53.129
32	29	22.076	55.301
33	32	21.077	60.961
34	27	20.147	63.292
35	16	19.280	63.850
36	18	18.471	63.862
37	21	17.713	64.472
38	16	17.003	64.531
39	12	16.337	65.683
40	24	15.711	70.056
41	17	15.122	70.290
42	19	14.566	71.639
43	14	14.042	71.639
44	16	13.548	72.083
45	11	13.080	72.414
46	14	12.636	72.561
47	13	12.216	72.611

TABLE 9 (continued)

48	16	11.818	74.091
49	10	11.440	74.272
50	13	11.080	74.605
51 - 55	51	50.576	74.608
56 - 60	36	43.781	75.991
61 - 65	39	38.317	76.003
66 - 70	35	33.854	76.042
71 - 75	29	30.156	76.087
76 - 80	22	27.055	77.031
81 - 85	18	24.428	78.722
86 - 90	26	22.180	79.380
91 - 95	26	20.241	81.019
96 - 100	15	18.556	81.700
101 - 110	27	32.862	82.746
111 - 120	17	28.240	87.220
121 - 130	22	24.563	87.487
131 - 140	17	21.586	88.461
141 - 150	8	19.139	94.944
151 - 160	5	17.101	103.507
161 - 170	12	15.385	104.252
171 - 180	6	13.924	108.761
181 - 190	12	12.671	108.797
191 - 199	5	10.471	111.656

TABLE 9 (continued)

B. Frequency-rank segment - generalized Log-rank

a=0.4451 b=0.9191 c=0.1604 H2=43

Rank of Term	Number of Term Occurrences	
	Observed	Expected
1	2406	3139.953
2	3973	4151.597
3	5042	5055.371
4	6079	5872.079
5	7077	6617.032
6	7687	7301.826
7	8273	7935.456
8	8822	8525.039
9	9297	9076.302
10	9717	9593.925
11	10136	10081.777
12	10523	10543.098
13	10904	10980.626
14	11267	11396.694
15	11622	11793.310
16	11975	12172.212
17	12327	12534.914
18	12657	12882.745
19	12984	13216.878
20	13302	13538.349
21	13615	13848.084
22	13917	14146.909
23	14209	14435.568

TABLE 9 (continued)

24	14496	14714.728
25	14780	14984.995
26	15059	15246.919
27	15329	15500.999
28	15595	15747.691
29	15857	15987.414
30	16116	16220.548
31	16367	16447.449
32	16618	16668.439
33	16864	16883.819
34	17104	17093.867
35	17335	17298.841
36	17565	17498.980
37	17790	17694.506
38	18015	17885.628
39	18233	18072.540
40	18446	18255.423
41	18655	18434.447
42	18860	18609.772
43	19062	18781.547

TABLE 10

EXHAUSTIVITY OF INDEXING - Medlars

Shifted Negative Binomial - n=5.30 w=0.882

Number of Postings per Document	Number of Documents Observed	Expected	Cumulative Chisquare
1 - 15	19	21.719	0.340
16	5	4.896	0.343
17	4	5.481	0.743
18	5	6.059	0.928
19	10	6.622	2.650
20	5	7.165	3.305
21	7	7.680	3.365
22	8	8.163	3.368
23	10	8.609	3.593
24	9	9.015	3.593
25	14	9.378	5.870
26	9	9.697	5.920
27	12	9.969	6.334
28	3	10.195	11.412
29	7	10.376	12.511
30	12	10.510	12.722
31	16	10.601	15.471
32	11	10.649	15.483
33	14	10.657	16.531
34	12	10.627	16.709
35	11	10.560	16.727
36	10	10.460	16.747
37	13	10.330	17.437

TABLE 10 (continued)

38	12	10.172	17.766
39	4	9.989	21.356
40	12	9.783	21.859
41	5	9.558	24.032
42	10	9.316	24.083
43	9	9.059	24.083
44	14	8.791	27.170
45	8	8.513	27.200
46 - 50	30	38.197	28.959
51 - 55	39	30.862	31.105
56 - 60	19	24.047	32.164
61 - 65	22	18.174	32.970
66 - 70	9	13.384	34.406
71 - 75	9	9.638	34.448
76 - 80	4	6.808	35.606
81 - 144	18	14.247	36.595

TABLE 11

EXHAUSTIVITY OF INDEXING - Cranfield

Shifted Negative Binomial - n=5.51 w=0.905

Number of Postings per Document	Number of Documents Observed	Expected	Cumulative Chisquare
1 - 20	16	18.501	0.338
21	4	3.291	0.491
22	5	3.618	1.019
23	5	3.946	1.300
24	5	4.272	1.424
25	2	4.594	2.889
26	7	4.908	3.780
27	7	5.213	4.393
28	8	5.506	5.522
29	4	5.787	6.074
30	5	6.052	6.257
31	5	6.302	6.526
32	8	6.534	6.855
33	4	6.747	7.973
34	6	6.942	8.101
35	13	7.116	12.965
36	10	7.271	13.989
37	6	7.406	14.256
38	5	7.520	15.101
39	8	7.615	15.120
40	3	7.689	17.980
41	5	7.744	18.952
42	8	7.780	18.959

TABLE 11 (continued)

43	3	7.798	21.911
44	6	7.799	22.326
45	12	7.782	24.612
46	6	7.750	25.007
47	11	7.702	26.419
48	12	7.640	28.907
49	6	7.565	29.231
50	1	7.477	34.842
51	10	7.378	35.773
52	10	7.269	36.800
53	8	7.150	36.901
54	9	7.022	37.458
55	5	6.886	37.975
56	14	6.744	45.782
57	3	6.596	47.742
58	5	6.442	48.065
59	6	6.284	48.078
60	11	6.123	51.963
61 - 65	20	28.111	54.303
66 - 70	23	23.895	54.337
71 - 75	27	19.844	56.917
76 - 80	22	16.149	59.037
81 - 85	11	12.910	59.320
86 - 90	6	10.159	61.022
91 - 100	14	13.923	61.023
101 - 164	14	17.033	61.563

TABLE 12

EXHAUSTIVITY OF INDEXING - CIJE

Shifted Binomial - p=0.295 N=25

Number of Postings per Document	Number of Documents Observed	Expected
1	1.	1.949
2	15.	20.393
3	138.	102.400
4	362.	328.502
5	820.	756.020
6	1401.	1328.666
7	1870.	1853.221
8	2093.	2104.824
9	2021.	1981.669
10	1529.	1566.284
11	832.	1048.633
12	483.	598.349
13	262.	292.102
14	159.	122.227
15	78.	43.838
16	45.	13.452
17	32.	3.518
18	16.	0.779
19	3.	0.145
20	6.	0.022
21	1.	0.003

Chi-square = 540.6

TABLE 13

QUERY STATISTICS

Number of Terms in Query	Number of Queries Cranfield	Medlars
1	-	-
2	-	1
3	-	-
4	3	2
5	5	2
6	2	4
7	3	1
8	4	2
9	4	2
10	1	2
11	-	2
12	1	2
13	-	1
14	-	-
15	1	-
17	-	2
22	-	1
Mean number of Terms in Query	7.3	9.2
Variance	7.17	22.31
Variance/Mean	0.98	2.43

TABLE 14

DISTRIBUTION OF QUERY TERMS OVER DOCUMENTS
Medlars - NEGATIVE BINOMIAL

Number of Query Terms in Document	Number of Documents Relevant Real	Relevant Expected	Nonrelevant Real	Nonrelevant Expected
0	18	21.2	7435	8245.8
1	56	45.4	2296	1818.6
2	58	52.5	647	401.1
3	46	43.3	159	88.5
4	18	28.6	28	19.5
5	13	16.1	9	4.3
6	8	8.0	5	0.9
7-10	4	5.9	0	0.3
Negative Binomial Parameters	w=0.1651, n=13		w=0.2206, n=1	

TABLE 15

DISTRIBUTION OF QUERY TERMS OVER DOCUMENTS
Cranfield - Negative Binomial

Number of Query Terms in Document	Number of Documents: Relevant Real	Relevant Expected	Nonrelevant Real	Nonrelevant Expected
0	16	15.6	4524	4066.2
1	28	38.4	2858	2938.1
2	41	49.2	1409	1592.2
3	45	43.7	668	766.4
4	37	30.2	299	346.4
5	16	17.3	124	150.2
6	21	8.5	61	63.3
7	2	3.7	21	26.1
8	2	1.5	3	10.6
9	1	0.5	0	4.3
Negative Binomial Parameters	w=0.0986, n=25		w=0.3613, n=2	

TABLE 16

KOLMOGOROV-SMIRNOV TEST FOR RECALL

Database	Calculated D Value	Critical Value	Result
Medlars	.40	.09	Reject H
Cranfield	.38	.09	Reject H

TABLE 17

DISTRIBUTION OF QUERY TERMS IN REAL AND SIMULATED
Medlars

Number of Query Terms in Document	Number of Documents			
	Relevant		Nonrelevant	
	Real	Simulated	Real	Simulated
0	18	16	7435	7580
1	56	34	2296	2470
2	58	25	647	370
3	46	8	159	31
4	18	2	28	4
5	13	0	9	0
6	8	0	5	0
7-10	4	0	0	0
Total	221	85	10579	10715

TABLE 18

DISTRIBUTION OF QUERY TERMS IN REAL AND SIMULATED
Cranfield

Number of Query Terms in Document	Number of Documents			
	Relevant		Nonrelevant	
	Real	Simulated	Real	Simulated
0	16	13	4524	4907
1	28	43	2858	3406
2	41	35	1409	1271
3	45	19	668	365
4	37	11	299	88
5	16	2	124	13
6	21	0	61	3
7	2	0	21	0
8	2	0	3	0
9	1	0	0	0
Total	209	123	9967	10053

TABLE 19

Term Co-occurrence Frequencies
Medlars

frequency	real	Simulated independent model	Simulated dependent model
1	293643	274170	278625
2	32158	27382	25003
3	9526	6619	6515
4	3942	2326	2286
5	1965	933	1035
6	1059	443	487
7	651	239	221
8	396	119	115
9	278	70	71
10	219	37	48
11-12	204	45	50
13-14	124	17	20
>15	199	13	10
TOTAL	344364	312413	314486
Chi-square compared to real		2000	2383

(pairs of columns were considered as a contingency table)

TABLE 20

Term Co-occurrence Frequencies
Cranfield

frequency	real	Simulated independent model	simulated dependent model
1	222229	216711	212588
2	48163	49493	43882
3	20690	19716	18023
4	11296	10258	9768
5	6887	6195	5951
6	4493	3984	3885
7	3252	2684	2865
8	2431	2016	1979
9	1858	1577	1478
10	1400	1126	1182
11-12	2156	1568	1596
13-14	1478	976	1153
15-19	2056	1324	1532
20-24	940	501	719
25-29	565	217	327
30-39	530	121	284
40-49	212	26	88
>50	192	6	30
Total	330837	318499	307330
Chi-square compared to real		1360	625

(pairs of columns were considered as a contingency table)

TABLE 21

Term Co-occurrence Frequencies
SPI

frequency	real	Simulated independent model	simulated dependent model
1	11905	7241	7488
2	1042	379	281
3	224	66	49
4	53	21	25
5	46	10	9
6	28	8	5
>6	22	7	7
Chi-square compared to real		112.06	224.34

(pairs of columns were considered as a contingency table)

TABLE 22

Term Co-occurrence Frequencies
CIJE

frequency	real	Simulated independent model	simulated dependent model
1	2112	2499	2233
2	482	337	435
3	161	89	158
4	74	25	59
5	44	22	40
6	26	8	30
7	28	10	9
8	10	1	6
9	11	1	6
10-11	16	2	8
12-15	6	2	8
16-19	9	1	3
20-29	12	1	5
>29	7	2	2
Chi-square compared to real		267.1	48.7

Total sample of 3000 pairs from each set.

TABLE 23

DISTRIBUTION OF TERM CORRELATIONS

SPI SUBSET OF 242 DOCUMENTS

Correlation (Rounded Down to nearest hundredth)	Number of Pairs of Clusters
-0.21	1
-0.20	1
-0.16	1
-0.14	2
-0.13	5
-0.12	3
-0.11	7
-0.10	14
-0.09	32
-0.08	48
-0.07	114
-0.06	238
-0.05	408
-0.04	1164
-0.03	4286
-0.02	11930
-0.01	6501
0.00	1845
0.01	13
0.02	20
0.03	21
0.04	32

TABLE 23 (continued)

0.05	36
0.06	54
0.07	43
0.08	33
0.09	42
0.10	47
0.11	43
0.12	30
0.13	71
0.14	45
0.15	70
0.16	61
0.17	39
0.18	104
0.19	26
0.20	62
0.21	49
0.22	40
0.23	73
0.24	66
0.25	37
0.26	9
0.27	83
0.28	10
0.29	13
0.30	64

TABLE 23 (continued)

0.31	13
0.32	42
0.33	8
0.34	77
0.35	13
0.36	3
0.37	4
0.38	13
0.39	15
0.40	71
0.41	13
0.42	5
0.43	18
0.44	10
0.45	2
0.46	8
0.47	5
0.48	3
0.49	79
0.50	10
0.51	1
0.52	5
0.54	4
0.55	1
0.56	1
0.57	11

TABLE 23 (continued)

0.59	2
0.60	2
0.62	14
0.63	1
0.64	2
0.66	4
0.70	14
0.72	2
0.76	1
0.77	5
0.81	12
0.84	1
0.86	1
0.89	1
0.91	2
1.00	6

TABLE 24

DISTRIBUTION OF CLUSTER CORRELATIONS

AFTER FIRST ITERATION (74 CLUSTERS)

Correlation (Rounded Down to nearest hundredth)	Number of Pairs of Clusters
-0.21	1
-0.16	1
-0.15	1
-0.14	2
-0.12	5
-0.11	6
-0.10	10
-0.09	27
-0.08	41
-0.07	66
-0.06	103
-0.05	208
-0.04	367
-0.03	524
-0.02	342
-0.01	36
0.00	482
0.01	23
0.02	30
0.03	29
0.04	27
0.05	25

TABLE 24 (continued)

0.06	27
0.07	34
0.08	30
0.09	26
0.10	21
0.11	24
0.12	19
0.13	24
0.14	13
0.15	17
0.16	14
0.17	13
0.18	15
0.19	4
0.20	11
0.21	5
0.22	3
0.23	3
0.24	8
0.25	2
0.26	2
0.27	2
0.28	1
0.29	4
0.30	2
0.31	4

TABLE 24 (continued)

0.32	2
0.33	3
0.34	3
0.35	2
0.38	1
0.40	1
0.41	1
0.43	1
0.45	1
0.48	1
0.49	1

TABLE 25

DISTRIBUTION OF CLUSTER CORRELATIONS

AFTER SECOND ITERATION (27 CLUSTERS)

Correlation (Rounded Down to nearest hundredth)	Number of Pairs of Clusters
-0.22	1
-0.21	1
-0.19	1
-0.18	1
-0.17	1
-0.14	1
-0.12	5
-0.11	4
-0.10	7
-0.09	8
-0.08	7
-0.07	14
-0.06	15
-0.05	23
-0.04	30
-0.03	22
-0.02	18
-0.01	11
0.00	105
0.01	9
0.02	6
0.03	9

TABLE 25 (continued)

0.04	5
0.05	11
0.06	11
0.07	3
0.08	2
0.10	6
0.12	1
0.13	2
0.14	1
0.15	1
0.16	2
0.18	1
0.20	1
0.24	2
0.26	2
0.31	1

TABLE 26

CORRELATION BETWEEN CLUSTER SIZE AND AVERAGE TERM FREQUENCY

Cluster Number	Number of Terms in Cluster	Average Frequency of Terms in Cluster
1	15	3.33
2	17	7.82
3	19	4.74
4	16	4.13
5	11	6.00
6	14	13.29
7	9	5.33
8	3	2.00
9	12	3.75
10	7	2.57
11	4	3.50
12	5	3.60
13	10	4.90
14	6	2.83
15	9	3.22
16	5	5.20
17	18	7.50
18	3	2.33
19	4	2.75
20	5	4.00
21	6	4.33
22	10	3.80
23	10	3.40

TABLE 26 (continued)

24	2	2.00
25	11	5.45
26	7	2.57
27	1	2.00

Correlation = .59

TABLE 27

DISTRIBUTION OF THE CLUSTER SIZE

Size of Cluster	Number of Clusters	
	First Iteration	Second Iteration
I	6	1
2	27	1
3	15	2
4	8	2
5	11	3
6	2	2
7	5	2
8	-	-
9	-	2
10	-	3
11	-	2
12	-	1
13	-	1
14	-	1
15	-	1
16	-	1
17	-	1
18	-	1
19	-	1

TABLE 28

FIT OF TERM DISTRIBUTIONS - CIJE-2

A. Frequency-size segment - Mandelbrot-Zipf.

Parameters estimated from CIJE

a=12539.2 b=1.752 c=5.335 H1=199

Number of Term Occurrences (x)	Number of Terms with x Occurrences Observed	Expected	Cumulative Chi-square
1	436	481.142	4.235
2	331	372.180	8.792
3	281	297.514	9.708
4	231	243.946	10.395
5	203	204.109	10.401
6	150	173.615	13.613
7	143	149.712	13.914
8	159	130.601	20.090
9	123	115.060	20.638
10	105	102.239	20.712
11	110	91.527	24.441
12	85	82.478	24.518
13	78	74.759	24.658
14	77	68.117	25.817
15	85	62.358	34.038
16	71	57.328	37.299
17	62	52.907	38.862
18	51	48.999	38.944
19	50	45.526	39.383
20	48	42.424	40.116

TABLE 28 (continued)

21	53	39.642	44.617
22	43	37.137	45.543
23	35	34.871	45.543
24	35	32.815	45.689
25	32	30.943	45.725
26	29	29.234	45.727
27	51	27.669	65.401
28	26	26.231	65.403
29	29	24.907	66.076
30	29	23.685	67.268
31	36	22.555	75.283
32	24	21.507	75.572
33	19	20.534	75.686
34	29	19.628	80.161
35	24	18.784	81.610
36	29	17.995	88.340
37	22	17.257	89.644
38	30	16.565	100.540
39	24	15.916	104.646
40	21	15.306	106.765
41	10	14.732	108.285
42	9	14.191	110.183
43	10	13.681	111.174
44	19	13.199	113.724
45	11	12.743	113.962
46	9	12.311	114.853

TABLE 28 (continued)

47	9	11.902	115.560
48	12	11.514	115.581
49	10	11.145	115.698
50	11	10.794	115.702
51 - 55	65	49.273	120.722
56 - 60	44	42.653	120.764
61 - 65	28	37.330	123.096
66 - 70	29	32.981	123.577
71 - 75	23	29.379	124.962
76 - 80	23	26.358	125.389
81 - 85	27	23.798	125.820
86 - 90	15	21.609	127.841
91 - 95	24	19.719	128.770
96 - 100	13	18.077	130.197
101 - 110	22	32.016	133.330
111 - 120	19	27.512	135.963
121 - 130	16	23.930	138.591
131 - 140	8	21.029	146.664
141 - 150	12	18.646	149.033
151 - 160	6	16.661	155.854
161 - 170	7	14.988	160.112
171 - 180	7	13.566	163.289
181 - 190	6	12.344	166.550
191 - 199	3	10.202	171.634

TABLE 28 (continued)

B. Frequency-rank segment - generalized Log-rank

Parameters estimated from CIJE

a=0.4451 b=0.9191 c=0.1604 H2=43

Rank of Term	Number of Term Occurrences	
	Observed	Expected
1	2724	3179.015
2	4494	4202.760
3	5581	5117.346
4	6534	5943.823
5	7451	6697.687
6	8352	7390.673
7	8849	8031.883
8	9297	8628.518
9	9698	9186.376
10	10097	9710.190
11	10496	10203.878
12	10891	10670.717
13	11251	11113.479
14	11582	11534.524
15	11902	11935.884
16	12222	12319.318
17	12542	12686.359
18	12854	13038.352
19	13160	13376.481
20	13464	13701.798
21	13761	14015.237

TABLE 28 (continued)

22	14056	14317.638
23	14342	14609.749
24	14620	14892.248
25	14893	15165.749
26	15166	15430.806
27	15437	15687.925
28	15704	15937.568
29	15968	16180.158
30	16226	16416.082
31	16477	16645.696
32	16726	16869.330
33	16973	17087.287
34	17216	17299.848
35	17457	17507.273
36	17695	17709.806
37	17932	17907.671
38	18168	18101.080
39	18400	18290.227
40	18628	18475.298
41	18852	18656.463
42	19075	18833.885
43	19291	19007.716

TABLE 29

EXHAUSTIVITY OF INDEXING - CIJE-2

Parameters estimated from CIJE

Shifted Binomial - p=0.295 N=25

Number of Postings per Document	Number of Documents Observed	Expected
1	2	1.958
2	12	20.474
3	127	102.742
4	465	.401
5	1051	757.630
6	1534	1330.687
7	1915	1854.917
8	2031	2105.474
9	1887	1981.080
10	1539	1564.870
11	810	1047.051
12	397	597.085
13	210	291.308
14	108	121.821
15	50	43.666
16	16	13.391
17	9	3.500
18	2	0.775
19	1	0.144
20	2	0.022

END

21-09-82

FIN