Information Processing & Management Vol. 28, No. 1, pp. 121-133, 1992
Printed in Great Britain.

0306-4573/92 $5.00 + .00

APPLYING INFORMETRIC CHARACTERISTICS OF DATABASES TO IR SYSTEM FILE DESIGN, PART I: INFORMETRIC MODELS

DIETMAR WOLFRAM
School of Library and Information Science, P.O. Box 413, University of Wisconsin-Milwaukee,
Milwaukee, WI 53201, U.S.A.

Abstract—This study examines how informetric characteristics of information retrieval (IR) system databases can be used to help the systems designer decide what types of file structures would provide the best performance for a given type of information system environment. In this first of two papers, the development of appropriate models describing database contents, to be used later in a simulation study, are dealt with. Database characteristics for which data were collected include: the index term frequency distribution, the distribution of terms used per query, and the distribution of term frequency selections. A shifted generalized Waring distribution was found to provide the best fit for the index term distributions with the large data sets used. For the terms used per query, a shifted negative binomial was found to provide a reasonable fit. A complex relationship was observed for the term selection distribution data, for which the empirical distribution is used. As well, four other hypothetical term selection relationships are presented. With this information, a simulation study examining system performance under different informetric environments can be undertaken.

1. INTRODUCTION

Despite the development of faster computing and storage machinery, the efficient design and maintenance of information retrieval (IR) systems for optimal performance are still of paramount importance. As an example, CD-ROM technology has provided a compact information storage medium, and in only a few years has quickly gained wide acceptance in libraries and information centres. However, CD-ROM technology is comparatively slower as a retrieval method than traditional magnetic disk. The design of more efficient retrieval systems could help to reduce such a drawback. As well, with the proliferation of increasingly larger databases requiring a greater amount of searching, efficient designs again help to reduce search overhead.

Informetrics has traditionally found applications in such areas as library collection management and the development of science policy in government. However, it could also provide a valuable tool with which to design and maintain IR systems. Using knowledge of the quantitative properties of the information contained in such systems—for example, the distribution of index terms and the database growth rate—decisions could be made relating to system design features and maintenance.

The purpose of this two-part study is to compare, using simulation techniques, the retrieval and space requirements performance of several types of IR systems using different file structures and informetric patterns of database contents and use. In this first paper, the informetric characteristics of existing databases that describe system contents and use are presented. Using these characteristics as the basis for a hypothetical system, simulated performance results can be examined to see how these characteristics affect performance (Wolfram, 1992). Specifically, the effect of different index term distributions and usage behaviours on the retrieval time and space requirements are examined. Comparisons are made with several file structures used in automated IR systems. By varying in a factorial study the parameters relating to the term distributions and term selection distributions, the comparison is extended to a number of retrieval system situations. The ultimate goal of such a simulation study is to determine whether specific index term distributions and selection patterns favour use of one database file design over another for improved accessibility and better space economy.

2. PREVIOUS RESEARCH

A number of studies have already explored several aspects relating to informetric modelling of IR systems. These have ranged from simple descriptive statistics of the usage of data fields, for example (Williams & Shefner, 1976; Ayres & Yannakoudakis, 1979), to more complex model development of index term usage and occurrence patterns using informetric distributions, up to the development of complete system models.

Models of index term distributions in inverted file IR systems have received the greatest attention in the literature. In an early study, Griffiths (1975) examines the frequency of input of index terms, as well as the number of index terms assigned to a document based on index terms used in several commercial retrieval system databases. The study was to serve as a preliminary investigation towards the development of an overall simulation model.

Nelson and Tague (1985) provide a comprehensive study in which different mathematical distributions are fitted to empirical data sets from several databases, with respect to the distribution of index terms over documents, the distribution of exhaustivity, and the distribution of term co-occurrences. A shifted binomial, negative binomial, Zipf, Log-Rank, and a proposed Split Size-Rank model were tested. The comparatively poor fit of size frequency models for the tail end of observed data distributions led the authors to propose a split model that relies on a size frequency model for the low-frequency terms and a rank frequency model for the more frequently occurring terms.

In a supplementary study of term distribution modelling, Nelson (1989) examines several theoretical distributions to determine which may provide better fits to observed data, and the appropriateness behind each distribution's use. The generalized Waring and generalized inverse Gaussian Poisson distributions are shown to provide improved fits over variations of the traditional Zipfian distributions used in the past.

Other characteristics of IR system index terms have also received attention in the literature. In another study conducted by Nelson (1988), the correlation of term usage and term occurrence frequencies are examined for an online public access catalogue. A significant correlation (Pearson's $r = 0.808$) was found to exist between frequency of occurrence and use. Nelson (1983) has also examined the co-occurrence of index terms in user queries. It was shown that the independence assumption of index term usage (i.e., the use of one term in a query is independent of the use of another) is unrealistic. For a small percentage of index term pairs, a high association coefficient is observed, making at least a binary dependence model necessary.

The simulation approach to IR systems incorporating informetric characteristics has been undertaken in a number of studies. Simulations can serve one of two purposes, as outlined by Tague *et al.* (1981): (a) for the effective representation of bibliographic items and use, to better understand the processes involved; or (b) for performance evaluation, using access time of records to find more effective ways to retrieve bibliographic data.

Cooper (1973) provides one of the earliest attempts at fully simulating the overall information retrieval process and the parameters involved. The Cooper model consists of five parts: a thesaurus generator, document generator, query generator, search routines, and evaluation routines. The results of the simulation as a tool for modelling and evaluating the retrieval process were inconclusive. He felt that because of the lack of comprehensive knowledge of the overall process and the intricacies related to the interaction of different variables not modelled, the simulation model was still deficient as an evaluative tool.

Nelson (1982), in his doctoral dissertation, develops a general probabilistic model for an IR system defined by an 11-tuple of sets and functions, taking into account the terms over documents and queries, the distribution of exhaustivity over documents and over queries, the distribution of co-occurrences of terms, and the distribution of relevant and nonrelevant documents over the number of terms.

Tague *et al.* (1981) examine some of the difficulties with the overall simulation of retrieval systems. They present a formal model of a bibliographic retrieval system consisting of a quintuple: (1) a set of document descriptions, (2) a set of query descriptions, (3) a set of rankings for (1), (4) a retrieval function, and (5) an evaluation function. They perform a simple simulation and provide algorithms used to generate a set of document descriptions (distributions for index terms, co-occurrence of terms, terms over documents) and a set of queries.

Whereas many of the above studies employ informetric distributions to aid in the modelling of IR systems, few studies have explored the application of informetrics as an aid in the design and maintenance of IR systems. Houston and Wall (1964) in an early paper on the topic examine post-coordinate indices for term usage and draw attention to the application of knowledge of term distributions to automated retrieval systems design. Zunde and Slamecka (1967) provide an information theoretic approach to index construction by examining performance efficiency of index terms for information transmission. The optimum index term distribution for a system for maximum information transmission is compared to the observed distribution as an indicator of performance efficiency.

Tague (1988) and Fedorowicz (1981a, 1981b, 1982a, 1982b) demonstrate how the use of appropriate mathematical distributions may be used in file design to estimate space requirements for different components of a retrieval system, including the size of the index file and the postings file. Tague and Nicholls (1987) report that appropriate estimation of a Zipf size variable may be used to estimate the maximum size of a postings list expected in a postings file. Wolfram *et al.* (1990) explore the use of literature growth distributions to examine bibliographic database growth rates. Such knowledge would provide the systems manager with an indication of future disk space requirements and when additional indexers are required to index new entries. Brooks (1987) examines whether Bradford analysis of authorship dispersion can be used for database design by attempting to correlate Bradford multipliers with database space economy. No significant relationship was found.

Sampson and Bendell (1985) discuss how knowledge of the Zipfian distribution of index terms may be used in database performance modelling of secondary key indices. With knowledge of the index term distribution of a system, based on a random sample taken from the system, one can predict the minimum index lookup time for entries. Bennet (1975) investigates a similar idea and provides the same recommendation. He also examines a method for efficient storage of citation numbers for the more frequently occurring terms, thereby utilizing less space.

The lack of any substantial literature that examines the use of informetrics in pragmatic terms, and in a comprehensive manner, as a tool for IR systems design and maintenance, provides many avenues of research that have not been examined extensively or have yet to be explored. It does not appear that factorial studies have been undertaken which take into account different file structures with different system parameters described by informetric distributions for improved accessibility and space savings. The preliminary model development for such a study is presented here.

3. DATA REQUIREMENTS AND COLLECTION

Prior to model development, it is necessary to collect appropriate data from which informetric models outlining the observed relationships can be fitted. A number of characteristics may be used to represent the functioning of an IR system. For the present study knowledge of three random components is necessary:

1. the distribution of terms used per query (Terms per Query Distribution) – to provide information relating to the size of user queries;
2. the frequency of occurrence of index terms in the database (Index Term Distribution) – to determine how terms are distributed within the database; and
3. the relationship between term frequency of occurrence and use (Term Selection Distribution) – to determine how terms are selected for use with respect to their frequency of occurrence in the database.

Different types of data can be stored in automated IR system databases. Thus obtaining data of several types provides a broader scope for the study. Examples of three types of systems were examined for data:

1. Online Public Access Catalogues (OPACs): OPACs contain catalogue data relating to the holdings of a library or information centre. Entries tend to consist of monographs and contain fields normally needed for cataloguing such materials.
2. Bibliographic Databases: Bibliographic databases provide a wider range of information than OPACs by including citations and possibly abstracts to periodical literature, conference proceedings, letters, reviews, and monograph data. Online databases available through vendors such as DIALOG and BRS as well as many CD-ROM databases are generally of this type.
3. Full Text Databases: In addition to providing bibliographic citations, such databases provide the text of the item being searched. Because they include passages of natural language text, their index term distributions are expected to differ from those not including natural language text. Such databases are also available through database vendors.

3.1 *Terms per query distributions*

In most online search systems, queries exist as one or more terms to be searched, strung together with boolean operators. Ideally, a transaction log of searches performed over a period of time could be collected and analyzed to provide the distribution of terms used per query. Of the three types of systems studied, data collection of this type was possible only for the OPAC. The OPAC system at the School of Library and Information Science (SLIS) at the University of Western Ontario (U.W.O.) at the time was used to collect query data for another research project underway at SLIS. The small number of searches collected over the spring term at SLIS by the project prevented sampling of the searches, so it became necessary to rely on all valid searches recorded.

Data for the bibliographic and full-text database systems were obtained from another source. A vast store of data extending back several years was available in the form of online search strategies for user search requests, performed by the online search librarians of the D.B. Weldon Library at U.W.O. It would not be unreasonable to assume that the wealth of search strategies are representative of the actual searches performed and of IR system searches in general. A random sample of 300 searches over the most recent two years provided an adequate sample for the bibliographic database with which to model the number of terms in a query.

Data representing searches from full-text systems were not as numerous as for the bibliographic searches. It was possible to extract such search data from the D.B. Weldon Library collection of searches, spanning 1987, 1988, and the first two months of 1989, but as with the OPAC data it was necessary to include all available searches without the benefit of sampling.

Two characteristics of the online searches performed at the D.B. Weldon Library were of concern: (1) The possible lack of inclusion of searches that use a small number of terms within the database of searches, and (2) The use of truncation symbols in search requests. In searches involving few terms, a written record of the search is occasionally not kept on file. The nature of such searches is generally informal, and thus can be excluded. Searches that employ truncation symbols may also cause difficulties in the identification of how many search terms were used, or were intended to be used, by the searcher. For the small proportion of such searches, the number of search terms searched was estimated from the likely terms to have been included in the index. When a doubtful estimation situation did arise, the search request in question was discarded and another was chosen.

Selection of mathematical distributions for model fitting was limited to Poisson-like functions. Very few queries existed with few terms or with many terms, but the majority of queries consist of a mid-range number of terms (Figs. 1 to 3; see the Appendix for the collected data). Probability functions, represented by $f(x)$, for two distributions were tested for goodness of fit to the observed data. These were:

(a) shifted Poisson

$$f(x) = \frac{e^{-\mu}\mu^{x-1}}{(x-1)!}, \quad x = 1,2,\ldots \tag{1}$$

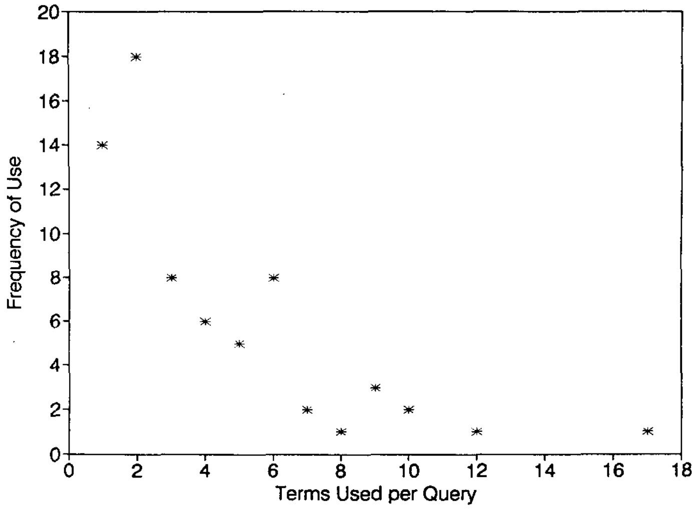

Fig. 1. Distribution of terms used per query—OPAC database.

where μ is a constant; and

(b) shifted negative binomial

$$f(x) = \binom{n+x-2}{x-2}(1-p)^n p^{x-1}, \quad x = 1,2,\ldots \tag{2}$$

where n and p are constants. Shifted forms of the distributions were used because of the nature of the data. The unshifted distributions begin with $x = 0$, whereas the terms per query data begin with $x = 1$.

A third distribution, the shifted binomial, was also initially considered. However, from the nature of the data, in terms of the relationship between observed means and variances, the calculation of parameter values was unstable and clearly inappropriate for representing the data.

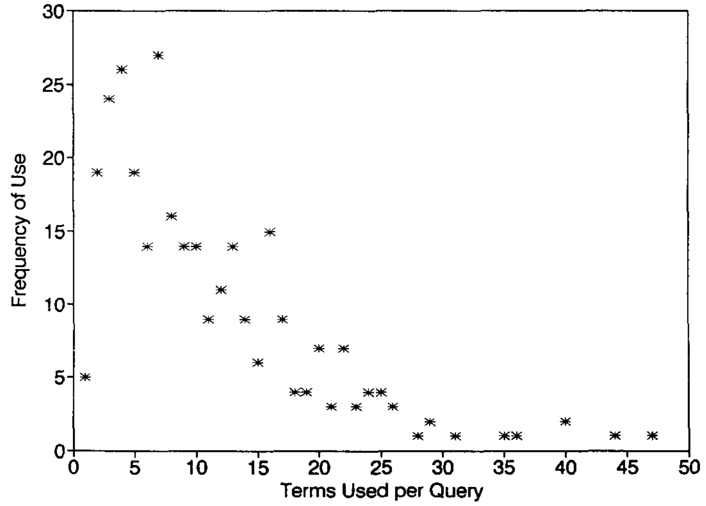

Fig. 2. Distribution of terms used per query—bibliographic database.

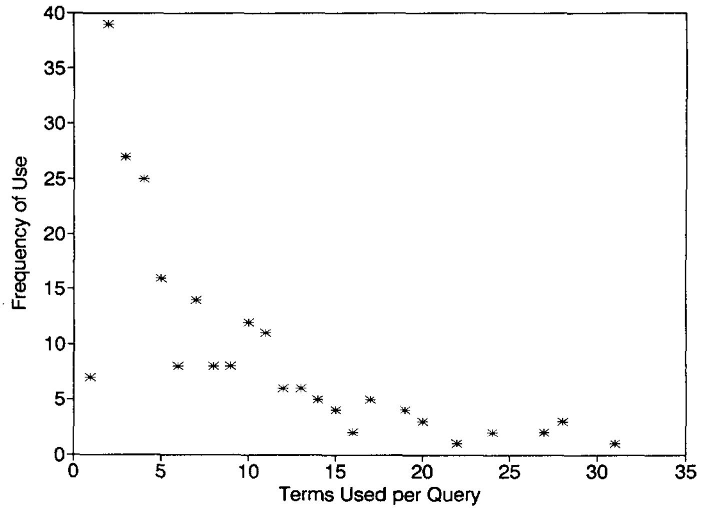

Fig. 3. Distribution of terms used per query—full-text database.

3.2 *Index term distributions*

Mathematical models for the distribution of the frequency of occurrence of index terms have been examined in several studies, reviewed earlier. For the current study, four distributions were selected for fitting to the collected data, either because of their traditional use in such research or because of their recently realized potential:

(a) Simple Zipf distribution: The proportion of index terms $f(x)$ appearing x times can be estimated by:

$$f(x) = a/x^b, \quad x = 1,2,\ldots, X\mathrm{max} \qquad (3)$$

where $a = 1/\sum_{x=1}^{X\mathrm{max}} (1/x^b)$ and b is a constant.

(b) Mandelbrot-Zipf: This is a generalized three parameter form of the simple Zipf:

$$f(x) = a/(x + c)^b, \quad x = 1,2,\ldots \qquad (4)$$

where a, b, and c are constants.

(c) Shifted negative binomial distribution:

$$f(x) = \binom{n + x - 2}{x - 2}(1 - p)^n p^{x-1}, \quad x = 1,2,\ldots \qquad (5)$$

where n and p are constants.

(d) Shifted generalized Waring distribution:

$$f(x) = \frac{\Gamma(v + \alpha)\Gamma(x + v - 1)\Gamma(x + \beta - 1)}{\mathrm{B}(\alpha,\beta)\Gamma(v)\Gamma(x + v + \alpha + \beta - 1)(x - 1)!}, \quad x = 1,2,\ldots \qquad (6)$$

where α, β and v are constants, $\Gamma(*)$ represents the Gamma function, and B(*,*) the Beta function.

Index term distribution data were obtained for the different types of IR systems, outlined in the previous section. The number of observations prevent a clear graphical presentation of the index term data. However, a summarization of the data appears in the Appendix. The SLIS OPAC database was used to provide the OPAC term distribution data, containing 39,617 records and 33,336 terms in the basic index. For the bibliographic and full-text databases, reasonably-sized databases from the DIALOG ONTAP selection were used. Data collection required that frequency of occurrence values be noted for each basic index term and then be tallied into the term distribution. For the bibliographic database, ONTAP Food Sciences and Technology Abstracts (file #251) was selected. The database represents six months of data (January 1985–June 1985) containing 10,156 records and 73,975 terms. For the full-text database, ONTAP Tax Notes Today was selected (file #199), containing six weeks (August 4, 1987–September 15, 1987) of full-text data, totalling 1,456 records and 35,089 terms. Unlike many of the previous studies, which have relied on relatively small data sets with hundreds of documents, this study relied on rather large databases containing many thousands of terms, thereby making model fitting even more difficult due to the large number of term sizes.

3.3 *Term selection distributions*

Because the relationship between frequency of term occurrence and frequency of use plays an important role in determining the probability of a term of a given size being used in a query, the ultimate system performance will be highly dependent on the type of relationship observed. Several studies reveal different opinions and findings on index term use and occurrence. Bennet (1975) indicates that no observable relationship exists, based on collected data. Salton (1975) hypothesizes that for best retrieval results, users should choose index terms of mid-range frequencies, as opposed to those that occur least and most frequently, because the mid-frequency terms provide the greatest discriminatory power. Nelson (1988) demonstrates that a significant correlation (Pearson's $r = 0.808$) exists between occurrence and use, with terms occurring most frequently being chosen most frequently in queries.

A variety of such relationships are conceivable, so several are proposed for this study. Data in the required format were very difficult to collect because few IR systems are specifically designed to tabulate usage values. It was therefore necessary to supplement the collected data with hypothetical distributions proposed by others for this activity. It was possible to collect data relating to term usage for the SLIS OPAC, as it keeps a count of the frequency of use for each term. The usage frequency for each term was extracted and then cumulated to form the term selection distribution. Four other hypothetical distributions are also considered and discussed below.

4. MODEL FITTING RESULTS AND DISCUSSION

The observed data collected were fitted to mathematical distributions to determine which distributions best modelled the data behavior. The technique for parameter estimation used was minimum chi-square. Minimum chi-square estimation, as Berkson (1980) has shown, provides good estimates for parameters and may even be regarded as more fundamental than more popularly used measures, such as maximum likelihood estimation. One main advantage of minimum chi-square is that it does not require the formulation of point estimators for model parameters, such as for method of moments and maximum likelihood estimation. For complex distributions, development of these estimators can be computationally quite intensive.

Minimum chi-square estimation attempts to determine most appropriate parameter values for distribution fitting by choosing those values that minimize the chi-square goodness-of-fit between the observed and fitted values, thus providing the best fit. Estimation of parameter values may be performed manually through iteration for distributions where only one or two parameters need to be estimated. The procedure may be automated for more complicated estimations by relying on minimization routines that attempt to bracket minimal estimates for each parameter. In this study the STEPIT minimization routine (Chandler, 1965) was used for estimation of the three parameter distributions.

Once parameter values have been estimated, the closeness of fit of the proposed distribution to the observed data may be compared using chi-square goodness-of-fit tests. Another method commonly used for this purpose is the Kolmogorov-Smirnoff goodness-of-fit test (Conover, 1980). Since this method usually assumes use of continuous data in cumulative form, not uncumulated discrete data as in this study, it was not used here.

Chi-square values were used for comparative purposes only. Models that produced the lowest value were chosen as providing the best fit. Since differences in chi-square values were so clearly defined, it was not necessary to take into account the effect of the slightly different numbers of degrees of freedom due to the number of parameters in each model. Models were also not tested in terms of level of significance for two reasons:

- Four of the seven data sets contain a large number of cells for chi-square fitting, even with appropriate collapsing of data values across cells. With such a large number of degrees of freedom involved, even small differences between observed and fitted values quickly accumulate, thereby producing a significant chi-square outcome.
- Most of the seven data sets represent populations, not samples from which inferences can be drawn. The use of inferential methods to draw conclusions would, thus, be inappropriate.

4.1 *Terms per query model fitting*

The nature of the collected data for the terms used per query hinted that distributions with Poisson-like characteristics should be selected for model fitting. The chi-square fittings for the three data sets employing the shifted Poisson distribution (Table 1) indicate a poor fit. The distribution contains only one parameter, and does not effectively model the range of behavior observed, particularly in the tail of the distribution, where the fitted values decreased much more rapidly than the observed values. The shifted negative binomial distribution provided much better fits for the data. The distribution contains two parameters and is more accommodating in modelling the observed behavior. For the bibliographic and full text databases the improvement is substantial.

4.2 *Term distribution model fitting*

Each of the term distribution data sets collected represented a large range of values to be fitted, due to the numerous terms. These values, in turn, represented a large number of distinct posting list lengths for the term distribution within each database. Traditionally, bibliometric distributions have been fitted to data sets where the maximal observed value and set of term occurrence values were comparatively small. For the collected data sets, the number of distinct term frequencies for each database was several hundred, with maximal values of 1,965, 8,451, and 16,369 for the full-text, OPAC, and bibliographic databases, respectively. With so many term frequencies, high values for goodness-of-fit results were expected. Despite this, with appropriate collapsing of data across term frequencies of occurrence at the tail end of distributions, reasonable fits were obtained for a few of the distributions tested.

The traditional single parameter Zipf model, because of its simple nature, provided a relatively poor fit, particularly for the bibliographic and full-text data (Table 2). However, the fit for the OPAC data was surprisingly good. The Mandelbrot generalization of the traditional Zipf distribution provided a noticeably better fit for all data sets. The three parameters allow the distribution to be more flexible in modelling the data. The shifted negative binomial distribution, despite being a two-parameter model, provided the poorest fit

Table 1. Terms per query data model fitting

Model system	Parameters		Chi-square	*df*
Shifted Poisson	μ			
Bibliographic	9.14		3412.89	17
Full-text	7.70		988.82	15
OPAC	3.74		36.25	7
Shifted negative binomial	n	p		
Bibliographic	1.76	0.15	29.95	23
Full-text	1.20	0.15	35.86	17
OPAC	1.28	0.30	6.38	6

Table 2. Term distribution model fitting

Model system	Parameters			Chi-square	*df*
Zipf	a[a]	b			
Bibliographic	0.564	1.881		6913.39	200
Full-text	0.416	1.551		1406.56	191
OPAC	0.543	1.827		304.45	124
Mandelbrot-Zipf	a	b	c		
Bibliographic	0.169	1.515	−0.599	755.26	198
Full-text	0.258	1.434	−0.352	704.79	189
OPAC	0.435	1.753	−0.134	233.31	122
Shifted negative binomial	n	p			
Bibliographic	0.074	0.003		9225.67	199
Full-text	0.129	0.004		4226.97	190
OPAC	0.121	0.007		7632.71	123
Shifted generalized Waring	α	β	v		
Bibliographic	0.671	0.153	3.959	448.37	198
Full-text	0.685	0.226	8.457	314.03	189
OPAC	0.856	0.358	2.270	185.24	122

[a] *a* is not estimated, but is dependent upon *b*.

of the four models tested. It was particularly deficient in attempting to fit the tail end of the data, producing high chi-square values.

Ajiferuke (1989) and Nelson (1989) have shown the generalized Waring distribution (Irwin, 1975a, 1975b) to be quite flexible for the purpose of modelling bibliometric data, and the fitting of these data confirms this finding. The model provided the best fit of the four distributions, performing considerably better than the traditionally used Zipf and Mandelbrot forms. The shifted generalized Waring provided the most difficulty for model fitting because of the complex form of the density function. Even without the necessity of formulating estimators for the distribution as in method of moments or maximum likelihood estimation, minimum chi-square estimation using an automated algorithm still resulted in a computationally intensive procedure, which required numerous iterations before a set of parameter values providing minimal chi-square results could be isolated.

4.3 *Term selection models*

Data relating to term use/occurrence relationships are generally not kept on IR systems and were largely unavailable for the study, limiting the range of observed distributions. It was, therefore, necessary to rely upon hypothetical distributions based on the statements of other researchers in previous studies. Raw data were initially collected from the SLIS OPAC identifying the frequency of use of each term. By cumulating the use of each index term by frequency of occurrence, it was possible to develop an empirical distribution which described how often terms occurring once were chosen, how often those occurring twice were chosen, etc.

The resulting distribution for the OPAC data, using a log-log transform for clarity, appears in Fig. 4. The cumulated data reveal that terms that occur few times are used most frequently, terms that occur most commonly in the database are used with mid-range frequency, and terms that occur with mid-frequency are, on average, used least frequently. The observed distribution can be described as being roughly parabolic. This goes contrary to Salton's hypothesis, where mid-range frequency terms should be used more frequently to provide the best retrieval results. The observed distribution may not be representative of all online searching, and is studied as only one of five proposed. It may be the case that this observed behaviour is a result of inexperienced student searchers, who most frequently rely on very broad or narrow terms for searching. The distribution, although far from irregular, proved to be too widely dispersed to fit adequately to a mathematical function, so it was decided to leave the data in the empirical format. By making the observed values pro-

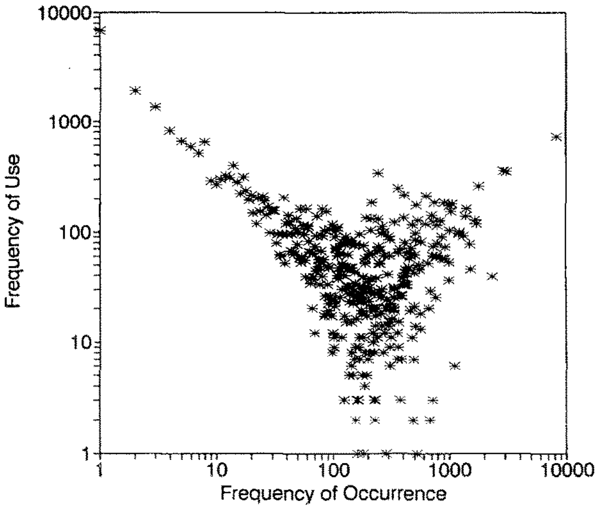

Fig. 4. OPAC term frequency of use versus frequency of occurrence.

portions for both frequency of occurrence and use, the empirical distribution may be generalized for use with different term distributions.

Because Bennet and others predict different relationships between use and occurrence, four other informetric relationships are also proposed for term selection, in addition to the observed relationship. The other four relationships were represented using hypothetical distributions. A graphical presentation of the four relationships appears in Fig. 5. These relationships include:

(a) Equal index term use probability: This behavior can be modelled by relying directly on the Zipfian relationship of the index term distribution. Overall, terms with only one posting will be chosen more frequently because they are more numerous.

(b) Equal use of each of the different frequencies of occurrence: This relationship assumes that each size of term has an equal probability of being used (i.e., a term that occurs once has the same probability of being used as a term occurring many times). It is represented by a uniform distribution over each frequency of occurrence.

(c) Direct relationship between frequency of occurrence and frequency of use: A weighting function that gives preference to high-frequency terms can be employed. The probability for use of a term of frequency x depends directly on the size of the term such that:

$$p(x) = \frac{x}{\sum_{i=1}^{X\text{max}} i} \quad x = 1,2,\ldots, X\text{max}. \tag{7}$$

(d) Mid-frequency terms are given preference: According to Salton's proposed use of index terms, the terms that occur in the mid-range are given a higher probability of use. This behavior may be modelled by employing a shifted negative binomial distribution, taking into account where the mid-range frequencies occur and the maximum observed values (Xmax). By assigning a desired probability of selection value, $f(x)$, at $x = 1$ and for $x = X$max giving:

$$f(1) = p^n \tag{8}$$

and

$$f(X\text{max}) = \binom{n + X\text{max} - 2}{X\text{max} - 2}(1-p)^n p^{x-1}, \tag{9}$$

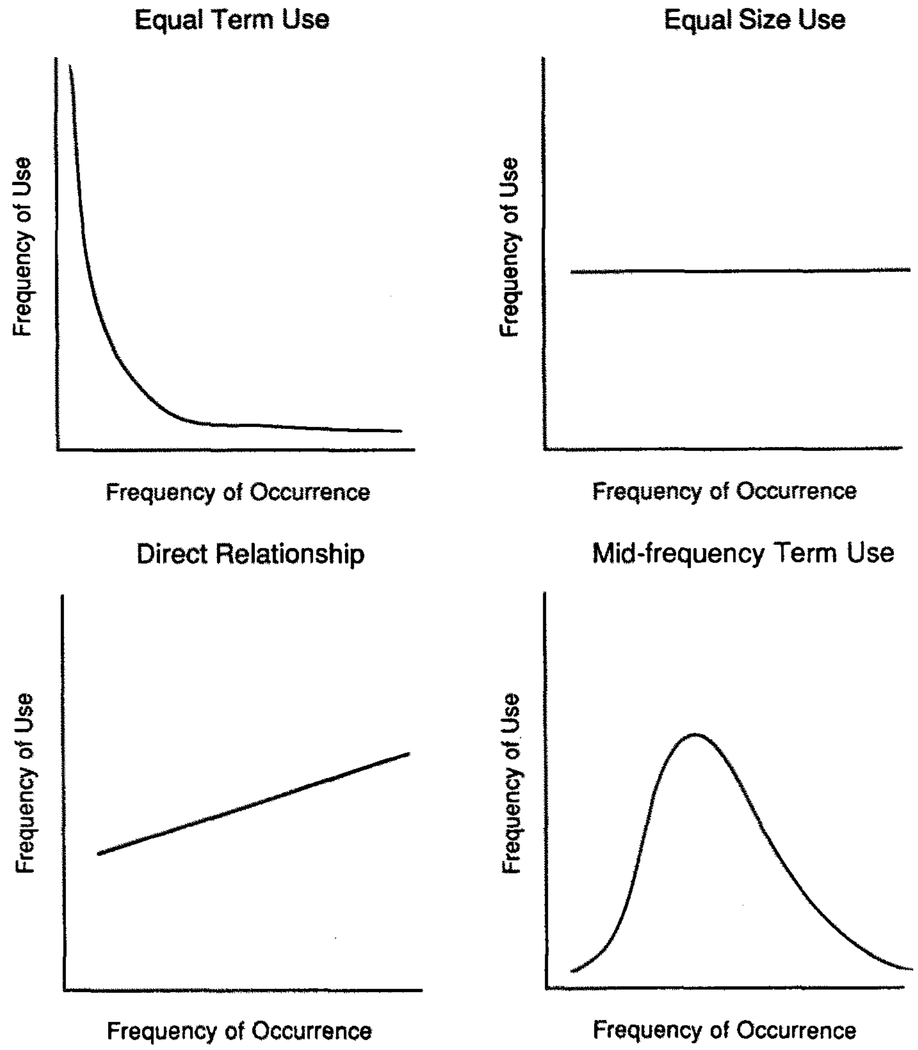

Fig. 5. Hypothetical term selection distributions.

one is left with two equations and two unknowns which may be solved using iterative methods. A range of parameter values that provide the desired behavior can then be used during the simulation to model this type of behavior.

Despite the fact that in four of the five cases the term selection distributions tested were hypothetical, based upon the findings and conjectures of others, the purpose here is to test system performance under such possible conditions and to determine what role the type of distribution played in determining retrieval time. Thus, the hypothetical nature of the term selection distributions in no way diverges from this overall purpose.

Based on these findings, an investigation of retrieval system performance may now be undertaken, using the best fitting models to represent the appropriate behaviors.

5. CONCLUSIONS

The present study has attempted to develop models of IR system database characteristics and use. It was found that for the large term distribution data sets, a shifted generalized Waring distribution provided the best fits. A shifted negative binomial distribution was found to provide reasonable fits in modelling the number of terms used per query. Preliminary data indicate that the relationship between term frequency of occurrence and use in a database is complex.

These characteristics of databases are areas that merit further investigation. Although index term distribution studies have already been undertaken in several papers, further examination is necessary. With the results of further studies, it may be possible to determine the range of existing distributions and whether similarities exist between database types, thus allowing for broader generalizations to be made. The ability to generalize these regularities could have benefits for deciding on design features and forecasting system performance. The changes in term distributions over time also need to be examined in greater detail.

Currently, all the data necessary to carry out such studies are not easily available. Term distribution data can be collected easily enough, but how often terms are used, and how queries are developed are aspects that are not usually stored. The extra space and computation necessary to accommodate such values within the indices may not make it cost effective.

In the second paper, the informetric models developed in this paper are used as the basis for an examination of system performance of different file structures under different informetric database characteristics.

Acknowledgements—The author would like to thank Drs. Michael Nelson, Mark Kinnucan, Sylvia Osborn, and Jean Tague-Sutcliffe for their helpful comments.

REFERENCES

Ajiferuke, I.S.Y. (1989). A probabilistic model for the distribution of authorships and a measure of the degree of research collaboration (Doctoral dissertation, University of Western Ontario, 1989). *Dissertation Abstracts International*, *50*, 03-A.

Ayres, F.H., & Yannakoudakis, E.J. (1979). The bibliographic record: An analysis of the size of its constituent parts. *Program*, *13*(3), 127–142.

Bennet, J.M. (1975). Storage design for information retrieval: Scarrott's conjecture and Zipf's Law. In E. Gelenbe & D. Poitier (Eds.), *International Computing Symposium* (pp. 233–237). Amsterdam: North Holland.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood. *The Annals of Statistics*, *8*(3), 457–487.

Brooks, T.A. (1987). Bradford analysis of authorship dispersion for database design. *Proceedings of the ASIS Annual Meeting*, *24*, 20–24.

Chandler, P.J. (1965). *Subroutine STEPIT: An algorithm that finds the values of the parameters which minimize a given continuous function.* Bloomington: Indiana University Quantum Chemistry Program Exchange.

Conover, W.J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: John Wiley & Sons.

Cooper, M.D. (1973). A simulation model of an information retrieval system. *Information Storage and Retrieval*, *9*, 13–32.

Fedorowicz, J. (1981a). Modelling an automatic bibliographic system: A Zipfian approach. (Doctoral dissertation, Carnegie-Mellon University, 1981). *Dissertation Abstracts International*, *42*, 03-A. (University Microfilms No. 81-18432).

Fedorowicz, J. (1981b). A Zipfian model of inverted file storage requirements. In W.G. Vogt & M.H. Mickle (Eds.), *Proceedings of the Twelfth Annual Pittsburgh Conference on Modelling and Simulation*, (pp. 1393–1399). Research Triangle Park, N.C.: Instrument Society of America.

Fedorowicz, J. (1982a). A Zipfian model of an automatic bibliographic system: An application to Medline. *Journal of the American Society for Information Science*, *33*, 223–232.

Fedorowicz, J. (1982b). The theoretical foundation of Zipf's law and its application to the bibliographic database environment. *Journal of the American Society for Information Science*, *33*, 285–293.

Griffiths, J.M. (1975). Index term input to IR systems. *Journal of Documentation*, *31*(3), 185–190.

Houston, N., & Wall, E. (1964). The distribution of term usage in manipulative indexes. *American Documentation*, *15*(2), 105–114.

Irwin, J.O. (1975a). The generalized Waring distribution, part I. *Journal of the Royal Statistical Society, Series A*. *138*(1), 18–31.

Irwin, J.O. (1975b). The generalized Waring distribution, part II. *Journal of the Royal Statistical Society, Series A*. *138*(2), 204–227.

Nelson, M.J. (1982). Probabilistic models for the simulation of bibliographic retrieval systems. *Dissertation Abstracts International* (Doctoral dissertation, University of Western Ontario, 1982).

Nelson, M.J. (1983). The use of term co-occurrence information in information retrieval. *Canadian Journal of Information Science*, *8*, 67–73.

Nelson, M.J. (1988). Correlation of term usage and term indexing frequencies. *Information Processing and Management*, *24*(5), 541–547.

Nelson, M.J. (1989). Stochastic models for the distribution of index terms. *Journal of Documentation*, *45*(3), 227–237.

Nelson, M.J., & Tague, J.M. (1985). Split size-rank models for the distribution of index terms. *Journal of the American Society for Information Science*, *36*, 283–296.

Salton, G. (1975). *A theory of indexing.* Philadelphia: Society for Industrial and Applied Mathematics.

Sampson, W.B., & Bendell, A. (1985). Rank order distributions and secondary key indexing. *Computer Journal*, *28*(3), 309–312.

Tague, J. (1988). What's the use of bibliometrics? In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88*, (pp. 271-278). Amsterdam: Elsevier Science Publishers.

Tague, J., & Nicholls, P. (1987). The maximal value of a Zipf size variable: Sampling properties and relationship to other parameters. *Information Processing and Management*, *23*(3), 155-170.

Tague, J., Nelson, M., & Wu, H. (1981). Problems in the simulation of bibliographic retrieval systems. In S.E. Robertson, C.J. van Rijsbergen, & P.W. Williams (Eds.), *Information Retrieval Research*, (pp. 236-255). London: Butterworths.

Williams, M.E., & Shefner, J. (1976). Data element statistics for the MARC II data base. *Journal of Library Automation*, *9*(2), 89-100.

Wolfram (1992). Applying informetric characteristics of databases to IR system file design, Part II: Simulation comparisons. *Information Processing and Management*, 28(1), 135-151.

Wolfram, D., Chu, C.M., & Lu, X. (1990). Growth of knowledge: Bibliometric analysis using online database data. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90*, (pp. 355-372). Amsterdam: Elsevier Science Publishers.

Zunde, P., & Slamecka, V. (1967). Distribution of indexing terms for maximum efficiency of information transmission. *American Documentation*, *18*, 104-108.

APPENDIX

Terms per query distribution data

Data appear as number of terms per query followed by frequency of occurrence.

OPAC database. 1:14, 2:18, 3:8, 4:6, 5:5, 6:8, 7:2, 8:1, 9:3, 10:2, 12:1, 17:1.

Bibliographic database. 1:5, 2:19, 3:24, 4:26, 5:19, 6:14, 7:27, 8:16, 9:14, 10:14, 11:9, 12:11, 13:14, 14:9, 15:6, 16:15, 17:9, 18:4, 19:4, 20:7, 21:3, 22:7, 23:3, 24:4, 25:4, 26:3, 28:1, 29:2, 31:1, 35:1, 36:1, 40:2, 44:1, 47:1.

Full-text database. 1:7, 2:39, 3:27, 4:25, 5:16, 6:8, 7:14, 8:8, 9:8, 10:12, 11:11, 12:6, 13:6, 14:5, 15:4, 16:2, 17:5, 19:4, 20:3, 22:1, 24:2, 27:2, 28:3, 31:1.

Term distribution data

Data appear as term postings list size followed by frequency of occurrence in the database.

OPAC database. 1:18758, 2:4661, 3:2234, 4:1321, 5:911, 6:648, 7:489, 8:414, 9:305, 10:215, 11:225, 12:183, 13:183, 14:168, 15:148, 16:126, 17:113, 18:94, 19:102, 20:79, 21:84, 22:62, 23:56, 24:66, 25:66, 26:53, 27:55, 28:54, 29:35, 30:47, 31:39, 32:38, 33:36, 34:22, 35:23, 36:27, 37:30, 38:30, 39:24, 40:36, 41:30, 42:31, 43:30, 44:22, 45:19, 46:20, 47:15, 48:18, 49:21, 50:18, 51-100:438, 101-250:271, 251-500:83, 501-1000:41, 1001-8451:19.

Bibliographic database. 1:49978, 2:6834, 3:3172, 4:1832, 5:1335, 6:946, 7:786, 8:647, 9:552, 10:469, 11:402, 12:333, 13:301, 14:281, 15:251, 16:201, 17:189, 18:196, 19:170, 20:164, 21:151, 22:112, 23:114, 24:112, 25:110, 26:106, 27:99, 28:101, 29:87, 30:81, 31:93, 32:88, 33:92, 34:75, 35:67, 36:72, 37:72, 38:61, 39:55, 40:56, 41:40, 42:54, 43:40, 44:41, 45:31, 46:38, 47:37, 48:48, 49:27, 50:53, 51-100:1134, 101-250:920, 251-500:350, 501-1000:187, 1001-16369:132.

Full-text database. 1:17132, 2:3798, 3:2057, 4:1407, 5:932, 6:689, 7:646, 8:546, 9:433, 10:345, 11:348, 12:301, 13:241, 14:252, 15:220, 16:208, 17:203, 18:176, 19:186, 20:137, 21:162, 22:135, 23:134, 24:118, 25:130, 26:85, 27:101, 28:87, 29:102, 30:81, 31:79, 32:69, 33:53, 34:56, 35:56, 36:60, 37:49, 38:60, 39:57, 40:55, 41:63, 42:40, 43:42, 44:39, 45:45, 46:47, 47:44, 48:35, 49:34, 50:35, 51-100:1090, 101-250:935, 251-500:397, 501-1000:200, 1001-1965:57.