APPLICATIONS OF INFORMETRICS TO INFORMATION RETRIEVAL RESEARCH

Dietmar Wolfram
University of Wisconsin--Milwaukee
dwolfram@uwm.edu

Abstract

A non-technical overview of two primary areas of study within the discipline of information science, information retrieval (IR) and informetrics, is presented. Informetric properties of IR systems as the basis for understanding IR system structure and generalizing human information seeking in electronic environments are discussed. Applications of informetric study of IR systems for more efficient and effective design and evaluation of IR systems are also presented.

Keywords: information science, information retrieval, informetrics

Introduction

Information science is an interdisciplinary field that encompasses the study of the production, organization, storage, retrieval, dissemination and use of information. Research may focus on the information user, the systems that provide access to information, or the interface between the two. Over the past fifty years a number of sub-fields have emerged within information science. Two primary areas of study within the discipline are information retrieval (IR) and informetrics. Each specialty has developed from different traditions, but have common areas of interest. In this paper, the author provides a non-technical overview of information retrieval and informetrics for the non-specialist, with a focus on the applications of the intersection of these two areas for IR system design and evaluation.

What is information retrieval?

Information retrieval is a selective process by which desired information is extracted from a store of information called a database (Meadow, 1992). Traditionally, IR systems have been used to locate text-based information, either the full-text of documents or document surrogates that summarize the contents of documents located outside of the database (e.g. bibliographic records). In recent years, information retrieval has broadened to include multimedia formats such as images, sound and video. IR system usage has also broadened during this time. Previously, information professionals were the primary users of IR systems, searching systems available through vendors such as DIALOG and EBSCO Information Services. The wider availability of online public access catalogues in libraries, CD-ROM database systems, and, most recently, web search engines, has made IR systems much more accessible to end users.

The process of interactive information retrieval involves a dialogue between the searcher and the IR system. The searcher initially submits a query to the IR system. Queries consist of one or more search terms and operators that define the parameters for records to be retrieved. The query terms are compared to an index of terms within the database using the operations (e.g. and, or, not) specified in the query. A list of records matching the query criteria is presented to the searcher for perusal. Based on the searcher's inspection of the records retrieved, the query may be reformulated. The process is then repeated.

On the surface, IR systems may resemble commonly used database management systems (DBMS). Although it is possible to develop an IR system using certain DBMS software, physical and philosophical differences distinguish these two types of systems. For example, the concept of *relevance* is central to information retrieval but does not play a role in DBMS interactions. Due to the ambiguities of language, not all items retrieved may be relevant to the searcher's information needs, despite having matched the query parameters. This is the challenge of IR: ensuring the timely retrieval of relevant items

while not retrieving those items that are non-relevant to the searcher's information need.

Numerous conceptual models have been developed for IR systems. Many of today's IR systems incorporate a Boolean approach where retrieval is based on an exact or partial match to a query. Many bibliographic database systems accessible within libraries or through database vendors such as DIALOG use this method. Also popular are probabilistic systems that take into account likelihood of relevance based on frequency of occurrence of search terms within documents, allowing retrieved items to be presented in rank order based on calculated relevance. Most World Wide Web search engines and other full-text IR systems rely on this approach. Still, other systems rely on a vector space model, where potential relevance is determined by proximity of documents to queries, represented as vectors in a multi-dimensional space (Salton & McGill, 1983).

Information retrieval remains a key research area within information science. Over the past twenty years, the study of information retrieval has expanded beyond efficiency and search mechanisms within the systems themselves to include human factors and searcher thought processes during IR sessions to better understand the mental processes user employ when performing information seeking tasks.

What is informetrics?

Informetrics is the quantitative study of information production, storage, retrieval, dissemination, and utilization. Informetric research investigates the existence of empirical regularities in these activities and attempts to develop mathematical models, and ultimately theories, to better understand information processes. The reader may also encounter two related, older terms--bibliometrics and scientometrics—that are often used synonymously with informetrics, or are considered to be sub-fields within informetrics (Brookes, 1988). Traditionally, bibliometrics has dealt with the study of print-based literatures (White & McCain, 1989) while scientometrics has focused on the statistical analysis of research patterns in the physical and life sciences (Diodato, 1994).

Major areas of study within informetrics include:

a) *Classic bibliometric' laws'* - These traditional areas of study deal with: Author productivity (Lotka, 1926), examining the publication contributions of authors to a given discipline; Journal productivity (Bradford, 1934), examining the concentration of articles in a subject area within a set of scholarly journals, and; Word usage (Zipf, 1949) examining the frequency of occurrence of words within texts.

b) *Citation and co-citation analysis* - This area looks at citing patterns of authors and publications or how authors are co-cited within articles, to determine strengths of relationships among authors, literatures or disciplines.

c) *Scientific indicators* - Studies examine the productivity of scientific output within disciplines or nations.

d) *Information growth and obsolescence* – This area investigates how literatures within subject areas grow over time.

e) *Document/information resource usage* – This area looks at how information resources are used over time.

Many informetric phenomena are characterized by inverse relationships within the processes studied. One often-studied example is the frequency of appearance of different words within a text. If one were to tally and then plot how many different words appear in this document once, twice, three times, etc., the resulting distribution of word frequencies would follow a type of inverse relationship, referred to as a 'Zipfian' distribution. This relationship is demonstrated in Figure 1.

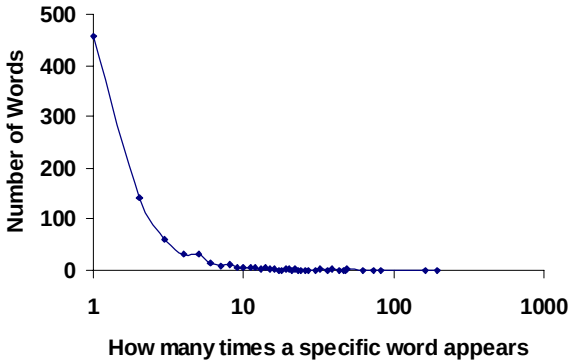

Figure 1: Frequency distribution of words appearing in this document

Note that many words appear only once and that few words appear many times. The ten most frequently appearing words in this text, excluding the references, title and abstract, are: the, of, to, and, a, IR, terms, in, for, system. They represent approximately 25% of all the words used in this text. Conversely, 448 of the 837 distinct words in this text (55%) appear only one time.

This relationship between the number of different items in the data studied and their frequency of appearance has been observed for many phenomena in recorded discourse. Knowledge of these empirical regularities may be used for the development of theory or applied to the development of information systems and services.

Specific applications of informetric research have included the development of national science policies, material selection and retention within libraries, and the measurement and the impact of scholarly journals within disciplines. Applications also extend to the IR environment, where knowledge of patterns of system contents and how they are used may be applied to IR system design and evaluation.

Informetric Study of IR Systems

To facilitate the retrieval process, documents added to an IR system must first be processed by parsing the document contents, usually down to the individual word level or even subword level. These words are then stored within term indexes. During the retrieval process in a simple Boolean IR system, the system will match the query components to a corresponding entry in the term index. The index entry will point to a list of document identifiers (also called a postings list) that contain the term. Postings lists are compared based on the search operations specified by the query. The final list is then presented to the user.

The structure of the IR system and system usage characteristics lend themselves well to informetric study. Researchers may examine how searchers interact with the system (IR system usage), or regularities in the content of the IR system databases and associated indexes (IR system content).

IR System Usage

Many aspects of system usage may be gathered for study to better understand how users interact with IR systems. These include:

a) *Terms used per query* - Searchers formulate queries by entering search terms. The topic of a query, its specificity, and the level of search experience of the user will often influence how many terms are used. When tallied, the terms used per query will frequently follow a unimodal distribution, where the most frequent number of terms will often be a small number, with far fewer queries using a large number of terms. Jansen et al. (in press), found that the average query length for searches conducted using the Excite search engine was 2.21 terms, but was very skewed with a mode of two terms.

b) *Distribution of query terms* - Just as with complete texts studied by Zipf, the frequency of use of specific words within queries follows an inverse relationship with many terms occurring only once, and few terms occurring many times (Jansen, Spink & Saracevic, in press).

c) *Query term co-occurrence* - Within multi-term queries, how frequently specific terms co-occur within queries also provides insight into searcher behavior. These data allow researchers to tackle questions such as 'Are combinations of specific terms frequently used within queries?' or 'Do searchers opt for combinations of frequently used and infrequently used search terms?' (Wolfram, in press)

d) *Searches conducted per user* - The number of searches a user will conduct also varies. Spink, Wolfram, Jansen & Saracevic (submitted) found that most users will only conduct a single search on a web search engine, while few people will conduct numerous searches.

e) *User browsing patterns* - Within systems containing hypertext linkages, researchers may study patterns in linkage traversal, and the types of linkages selected when searching for information (Huberman et al., 1998; Qiu, 1994).

IR System Content

The content of IR systems may similarly be studied for empirical regularities. Knowledge of these regularities may then be applied to the physical and logical design of IR systems for more effective and efficient access to system contents:

a) *Distribution of index terms* - How frequently specific terms appear within indexes has been studied by several researchers (Griffiths, 1975; Fedorowicz, 1982; Nelson & Tague, 1985; Nelson, 1989; Wolfram, 1992a). The inverse pattern observed by Zipf for lengthy texts is also observed for index terms; however, researchers have shown that more sophisticated models than those used by Zipf are needed to adequately fit the observed data.

b) *Term exhaustivity* - Subject headings or descriptors are often assigned to database records to provide additional terms by which a record may be retrieved. A unimodal distribution results, where more terms may be used for some records than for others (Nelson, 1982).

c) *Term co-occurrence* - The patterns of term co-occurrences within an IR system has also been shown to follow an inverse relationship (Nelson, 1983). Wolfram (1996) used descriptor term co-occurrence to develop a simulation model for representing inter-record linkage structure in a hypertext bibliographic retrieval system, where common occurrences of descriptor terms of records were used as the basis for inter-record linkages.

d) *Document citation and co-citations (or linkages)* - How one document cites another document, or how frequently two documents are co-cited by other documents may be

used to determine how closely related two documents may be. The same idea may also be applied to hypertext-based environments, where linkage relationships may be used to determine how closely related two web pages are.

e) *Database growth* - The growth of IR system contents will depend on the nature of the subject(s) included within the database. Some disciplines observe a slight exponential growth rate in their literatures, which would be reflected in the number of records included within the system. Growth of the World Wide Web, as an example of a very large information system, has recently been studied (Huberman & Adamic, 1999).

Applications to IR System Design and Evaluation

There are many applications of informetric studies within IR. To date, these applications have not been fully explored in IR system research. This is changing with the greater accessibility to large quantities of data related to the internal workings of IR systems and their use.

The use of basic informetric properties of text is implicit in some of today's IR systems. At least one search engine on the Web, Google (http://www.google.com), takes into account relationships between web documents based on hypertext linkages. Similarly, the frequency of occurrence of search terms within documents is often used as the basis for developing relevancy rankings within full text retrieval systems. Documents with many occurrences of the search terms will be presented to the user first, based on the assumption the more the search terms appear in a document, the more likely it is to be relevant to the user's information need. Algorithms for determining relevancy will also incorporate other features such as the document's length and the location of the terms within the document.

One of the primary applications of informetrics for IR system maintenance is space planning. Although electronic storage space costs have decreased over the years, the trend towards large databases containing many millions of records still makes this an important consideration. Tague (1988) and Fedorowicz (1981) demonstrated how the use of mathematical models could be used in file design to estimate space requirements for different components of an IR system, including the size of the index file and the postings file. Tague and Nicholls (1987) reported that appropriate estimation of a Zipf variable may be used to determine the maximum size of a postings list within a postings file. Such knowledge would provide the systems manager with an indication of future space requirements and when additional indexers may be needed to index new entries. Recently, Huberman & Adamic (1999) examined models for the growth of the World Wide Web and its implications for finding information on the Web.

Sampson and Bendell (1985) discussed how knowledge of the Zipfian distribution of index terms could be used in database performance modeling of secondary (non-unique) indexes. With knowledge of the index term distribution of a database, one can predict the minimum index lookup time for entries. If the observed performance is found to be sub-optimal, this could indicate that index term entries are not arranged on disk to require the fewest number of lookups. One may also use the Zipfian distribution of terms to determine which terms will be considered 'stopwords.' A stopword is a term that will not be indexed or processed in a query because of its high frequency. It is deemed to be of little value in the search process. For example, in most text-based IR systems, articles such as 'the' occur too frequently to be of value. These words provide no context or topicality for a search. The same may be true of frequently occurring nouns in subject-oriented databases, e.g. excluding a word such as 'science' in a science-oriented database since every record in the database probably deals with science. Salton and McGill (1983) hypothesized that the best search terms to use were those that occurred with mid-range frequency within the database, since they would be more discriminating than frequently occurring terms, but would cast a broader net than infrequently occurring terms.

Once the separate components of an IR system have been effectively modeled, one can integrate the parts into a comprehensive model. The resulting system model may then be used to emulate a hypothetical IR system using simulation techniques. Simulations serve two purposes: 1) They can be used for the effective representation of the IR system to better understand the underlying processes (Nelson, 1982); 2) They may be used for performance evaluation to find more efficient ways to develop the system. Wolfram (1992a, 1992b) developed a simulation model to test different file structures for accessing records in information retrieval systems. The model incorporated distributions for terms used per query, index term distributions, and distributions for index term selection. By varying the parameter values for these distributions, the author was able to test the performance of various file structures under different database and search characteristic environments. From this, recommendations could be made for the use of specific file structures under different circumstances.

More recently, the popularity of the Internet, and particularly World Wide Web, has prompted informetric studies of Internet contents and usage. This area of informetrics is sometimes referred to as 'cybermetrics' (see for example, http://www.cindoc.csic.es/cybermetrics/cybermetrics.html). This research also has implications for web-based IR system

design. Lawrence and Giles (1998) examined coverage of various search engines, and concluded that, at the time of the study, no single search engine indexed more than about one-third of the web. Wolfram (in press) looked that the co-occurrence of query terms submitted to the Excite search engine and concluded that pre-processing of postings lists for frequently used pairs of terms would not decrease search times noticeably due to the relatively small number of pairs of terms that co-occurred quite frequently. The author did conclude, however, that locating postings for the most frequently used individual terms in a faster area of computer memory was worth investigating since the 10 most frequently used terms, representing 0.01% of all terms, constituted 5% of all terms used. Patterns of query term use and their co-occurrence could also have applications for marketing purposes on web sites, where advertising in related areas may be selected based on these relationships.

Summary and Further Research

Although both information retrieval and informetrics research have received much attention in the information science literature, for the most part they have been investigated separately. This has changed with the wider availability of IR systems and the ability to capture data on system contents and usage. To date, however, the study of informetric properties of IR system contents and their usage have not been widely explored.

With new IR system technologies now indexing terabytes of information, a better understanding of the effects of increased database size on system structures and the implications for searchability is needed. As more documents are indexed within a system, the growth of term indexes will likely slow since most indexable terms will have already been incorporated in the index. The postings file, however, will continue to grow, with the number of uniquely occurring terms decreasing proportionately compared to the number of non-unique terms. Questions of how this will affect the processing of queries arise. More effective search and retrieval methods will be needed to allow users to quickly find information of interest, instead of becoming overwhelmed by too much information.

User studies of IR systems to understand the range of behaviors, and typical search behaviors and interactions, require further investigation. In the past, it was more difficult to obtain large quantities of data to study empirical regularities of query formulation and system usage. Most IR user studies have involved 'canned' sessions where small numbers of atypical users have been studied. Today, these data for end users are more easily obtained. Usage characteristics that merit closer investigation include changes in user search habits over time (e.g. numbers of searches, topics, most frequently used search terms, use of more advanced search features, numbers of search terms within queries), error analysis in query formulation, and the search engine/database selection process undertaken by searchers. With a better understanding of general search behaviors of users and database characteristics, recommendations may be made for system design features for improved access and performance.

References

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, 137, 85-96.

Brookes, B. C. (1988). Comments on the scope of bibliometrics. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88*, (pp. 29-40). New York: Elsevier Science Publishers.

Fedorowicz, J. (1981). Modeling an automatic bibliographic system: A Zipfian approach. *Dissertation Abstracts International*, 42, 03-A.

Fedorowicz, J. (1982). The theoretical foundation of Zipf's law and its applications 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.

Huberman, B. A. & Adamic, L. A. (1999). Growth dynamics of the World Wide Web. *Nature*, 401, 131-133.

Huberman, B. A., Pirolli, P., Pitkow, J. E., & Lukose, R. M. (1998). Strong regularities in World Wide Web surfing. Science, 280 (April 3): 95-97.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, 16(12), 317-323.

Jansen, B. J., Spink, A., & Saracevic, T. (in press). Real life, real users, and real needs: A study of user queries on the web. *Information Processing and Management*.

Lawrence, S., & Giles, L. (1998). Searching the World Wide Web. *Science*, 280 (April 3), 98-100.

Meadow, C. T. (1992). *Text Information Retrieval Systems*. San Diego: Academic Press.

Nelson, M. J. (1982). Probabilistic Models for the Simulation of Bibliographic Retrieval Systems. Unpublished doctoral dissertation, University of Western Ontario, London, Canada.

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

Qiu, L. (1994). Frequency distributions of hypertext path patterns: A pragmatic approach. *Information Processing and Management*, 30(1), 131-140.

Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill Books Co.

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

Spink, A., Wolfram, D., Jansen, B. J. & Saracevic, T. (submitted). Searching the Web: The public and their queries.

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

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

White, H. D. & McCain, K. W. (1989). Bibliometrics. in M. E. Williams, (Ed.) *Annual Review of Information Science and Technology*, (pp. 119-186). New York: Elsevier Science Publishers.

Wolfram, D. (1992a). Applying informetric characteristics of databases to IR system file design, Part I: Informetric models. *Information Processing and Management*. 28(1), 121-133.

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

Wolfram, D. (1996). Inter-record linkage structure in a hypertext bibliographic retrieval system. *Journal of the American Society for Information Science*. 46(10), 765-774.

Wolfram, D. (in press). Term co-occurrence in Internet search engine queries: An analysis of the Excite data set. *Canadian Journal of Information and Library Science*.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.