The Impact of Term-Indexing Characteristics on a Document Space

L'impact des caractéristiques des termes d'indexation sur un espace documentaire

Dietmar Wolfram and Jin Zhang
School of Information Studies
University of Wisconsin-Milwaukee
Milwaukee, WI
dwolfram@uwm.edu
jzhang@uwm.edu

Abstract: The authors investigate the impact of term frequencies, term distributions, and indexing exhaustivity on the representation of document spaces in a visual, vector-based retrieval environment. Using actual and simulated document sets, the authors compare document-space densities resulting from combinations of indexing characteristics and inclusion or exclusion of index terms of a given frequency. Singly occurring terms are demonstrated to contribute significantly to defining the document space density, which has implications for the retrieval of documents. Indexing exhaustivity also plays an important role in shaping the document space, with higher exhaustivities resulting in the lowest document-space densities. The implications for automatic indexing in IR systems are discussed.

Résumé : Les auteurs étudient l'impact de la fréquence et de la distribution des termes, et de l'exhaustivité d'indexation sur la représentation des espaces documentaires dans un environnement de repérage visuel vectoriel. En utilisant des jeux de documents réels et simulés, les auteurs comparent les densités des espaces documentaires résultant des combinaisons des caractéristiques des termes, et de l'inclusion ou de l'exclusion des termes d'indexation pour une fréquence donnée. Il est démontré que les termes à occurrence unique contribuent de manière significative à définir la densité de l'espace documentaire, ce qui a des implications pour le repérage de documents. L'exhaustivité d'indexation joue également un rôle important dans la définition de l'espace documentaire : de plus hauts niveaux d'exhaustivité résultent en de plus basses densités d'espace documentaire. Les implications pour l'indexation automatique dans les systèmes de RI sont présentées.

Introduction

The performance of an information retrieval (IR) system can be measured

from different perspectives. What is important for the user isn't necessarily the same as for the system designer. Ultimately, how well a system performs is determined by its ability to discriminate between the indexed documents, or their representations as bibliographic records, so that only the most relevant documents are retrieved. Therefore, in any good IR system, those documents that are deemed most relevant to a user's query should be distinguishable from those that are less relevant. This is achieved through various indexing and retrieval techniques.

Recent developments in information retrieval system design have made it possible to represent the interrelationships among documents using information visualization techniques. Visualization affords the unique opportunity of summarizing complex relationships in a comprehensible and visual manner for the user—something that is not possible when presented with a single value or linear list of ranked documents from a traditional IR system. The purpose of the present study is to explore the influence of system-indexing characteristics on the resulting document space in a vector-based, visual IR environment, and to demonstrate a methodology for visualizing and comparing document environments. It is widely recognized that the ability to distinguish documents from one another is a key determinant in retrieval effectiveness (Salton 1989; Korfhage 1997). Visualization makes it possible to characterize a document space by defining differences between documents as angles or distances. Because document characteristics are determined by the indexed content, it not unreasonable to conclude that the indexing characteristics of the document set will impact the structure of the document space and the relationships of documents to one another.

Several research questions guide the present research. First, how does the presence or absence of terms of different frequencies impact the document clusters of a document space? With known Zipfian distributions of term frequencies, where many terms occur only one time, are they important in defining the document set and characterizing documents? Second, what is the impact of postings (term "tokens") versus distinct terms (term "types") on the document space? Third, how do the indexing characteristics of different term distributions and indexing exhaustivities (terms assigned per document) impact the document space? This research has implications for automatic indexing within IR systems, where the most desirable environment for retrieval is one in which documents are most easily distinguished from one another.

Previous research and underlying concepts

Information visualization simplifies information representation by applying visual display processing to abstract information. Information visualization combines aspects of information behaviour, human-computer interaction, data mining, imaging, and graphics. It focuses on information that is often abstract. Recently, many visual information models or systems have emerged. These systems provide visual information environments for users to browse and interact with information (Benford et al. 1995; Kim and Korfhage, 1994; Olsen and Korfhage, 1994), or visually demonstrate attributes of documents or semantic relationships (Fekete and Dufournaud 2000; Shneiderman et al. 2000; Helfman 1994), or permit visualization of internal mechanisms for retrieval processing (Zhang 2001; Zhang and Korfhage 1999; Nuchprayoon and Korfhage 1994; Young and Shneiderman 1993). However, the application of visualization techniques in the information sciences, in addition to its applications in information seeking and searching, may also shed light on information analysis. For instance, visualization can be applied to term discriminative capacity analysis (Zhang and Wolfram 2001), document classification analysis (Liu et al. 2000), citation analysis (Small 1999), and full-text analysis (Fekete and Dufournaud 2000; Hearst 1995; Helfman 1994).

Experiments with IR system contents have examined the role of system characteristics on retrieval effectiveness or system storage requirements. For example, Burnett et al. (1979), examined the effect of the size of controlled index term vocabularies on retrieval performance using the Cranfield test collections. They found retrieval performance improved with an increase in the number of terms used to index documents, but the rate of improvement decreased as the number of index terms approached the complete set of terms from the original set. In a related study, Willett (1979) examined the use of fixed-length character strings for controlling the size of an indexing vocabulary on the same Cranfield data sets using hashing, truncation, and n-gram encoding techniques.

Empirical regularities of IR system content have been studied since the earliest days of computerized IR. Among the most widely studied aspect of IR system content has been the frequency distribution of index terms. The observed inverse relationship is typical of many informetric processes, where a small proportion of terms occur with great frequency and a large proportion of terms occur only once or twice. Researchers have

attempted to model the observed term distribution behaviour by fitting observed data sets to theoretical distributions that best model the observed patterns. Most studies to date have focused on "Zipfian" models. A simple size-frequency Zipf model (Tague and Nicholls 1987) can be characterized as:

$$f(x) = \frac{a}{x^b} \quad x = 1,2,3,...x_{max} \qquad (1)$$

where $f(x)$ corresponds to the proportion of all distinct terms (or term types) that occur x times, with a and b representing parameters to be fitted to observed data sets.

Generalizations of the Zipf distribution, such as a three-parameter Mandelbrot-Zipf have been shown to provide better fits than the traditional Zipf (Wolfram 1992). Similarly, Fedorowicz (1982) relied on different formulations of Zipf's law to model the distribution of terms in the MEDLINE database. Nelson (1989) fitted a generalized Waring and generalized inverse Gaussian-Poisson distribution to six data sets, demonstrating the feasibility of using more sophisticated models for fitting index term distributions. However, these models are less tractable than the more straightforward Zipf models. Models based on inverse power laws, although not always providing the best fits, lend themselves to experimental studies examining changes in term distribution characteristics. In the case of a Zipf distribution, different values in the a and b parameters reflect changes in the percentage infrequently occurring and frequently occurring terms respectively, thereby permitting manipulation of the shape of the distribution.

Another important feature of IR systems in modelling system content is the document indexing exhaustivity, or the number of index terms assigned to a document. Fewer models have been proposed for fitting exhaustivity distributions. Bird (1974) used Poisson, binomial, and negative binomial distributions, with some success for each depending on the nature of the distribution based on the relationship between the mean and variance of the data sets. Since then, researchers such as Nelson and Tague (1985) and Wolfram (1992) have also applied shifted forms of the Poisson, binomial, and/or negative binomial distribution with some success due to the "lumpiness" of the observed data sets. However, for general modelling purposes, such as computer simulations of IR systems where means and variances are known, the negative binomial distribu-

tion is the preferred model for its flexibility. One form of the negative binomial distribution takes the form

$$p(x) = \binom{\kappa + x - 1}{\kappa - 1} \pi^{\kappa} (1-\pi)^{x} \quad x = 0, 1, 2, \ldots \quad (2)$$

where p(x) represents the probability of a document containing x terms, with parameters ($0 < \pi < 1$) and $\kappa > 0$.

The present study combines aspects of computer modelling and simulation with visualization to investigate the impact of term types and tokens, in addition to different indexing characteristics, on the document space defined by a set of indexed documents. This research builds on another study (Wolfram and Zhang 2002) examining the impact of term distributions and indexing exhaustivity on a document space. The present study shifts the focus to the actual frequencies of terms and the influence of term types and tokens.

Method

Descriptor data for 500 bibliographic records dealing with library and information science were extracted from the NTIS database. The data represented 1,641 distinct terms (types) and 4,863 postings (tokens). Although small when compared with today's gigabyte IR studies, the authors wished to keep the data set size manageable for processing purposes within the visual environment. Also, the data set can be thought of as a subset of a much larger database, used as the starting point of a more detailed search. In visualization research, the computational burden associated with visualizing thousands of records simultaneously, along with the cognitive overload for users, makes the processing of large data sets impractical computationally and overwhelming for the user.

Descriptor frequency and indexing exhaustivity data for the records were stored in a Microsoft Access™ database. The use of a structured database environment permitted easy manipulation of the descriptor sets and made it possible to convert the descriptor data to vector representations incorporating inverse document frequency (*idf*) term weights. Tokens for each term type were assigned the same term weight because intra-document term weights for each term were not available. The resulting data sets were then imported into the distance-angle-based visual tool called

DARE (Distance Angle Retrieval Environment) (Zhang 2000; Zhang and Korfhage 1997). *DARE* is a two-dimensional visual retrieval tool that presents a graphical representation of the visual distance and the visual angle in a vector-space environment as *X*-axis and *Y*-axis values respectively.

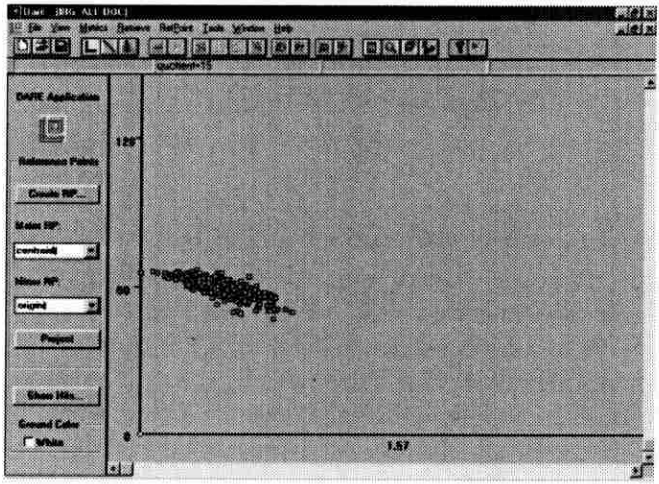
Construction of a visual space is one of the fundamental parts of a visual information tool, where objects are displayed and manipulated. Information attributes are usually employed to construct a visual space for a visual information tool. Two attributes of an object, visual distance and the visual angle, are defined in the *DARE* environment as the *X*-axis and *Y*-axis for the visual space respectively. The algorithm used in *DARE* requires users to define two reference points in a vector-based document space. The reference points reflect users' information needs or current search interest. These two reference points are used as the reference system to project all documents in the document space onto the *DARE* visual space. These two reference points then can determine a line in the document vector space.

Observe that the two reference points and any document can determine a hyperplane in the document space. Within the plane, the distance between the document and one of the two reference points, and the angle formed by the two lines (one is determined by the document and the selected reference point and the other by the two reference point) can be measured. The distance and angle of the document are defined as the visual distance and visual angle, respectively. These two parameters for a document are always available no matter how high the dimensionality of a document vector space is and where the document is located if the two reference points are clearly defined.

The two parameters are crucial and fundamental for the visual-space construction. Based on the two parameters of a document, the document can be easily projected onto a two-dimensional visual space. The visual space is limited to an open rectangular area. The *X*-axis of the visual area ranges from zero to π radians due to the symmetrical characteristic against the reference line determined by the two reference points in measuring a visual angle of a document. Since there is no limitation on a visual distance of a document, the *Y*-axis can range from zero to infinity, theoretically. The document space for a document set can be represented as a two-dimensional scatter plot of the distance and angle measure of each document in relation to the document-set centroid (i.e., the

representation of the "average" document within the space). A sample output from the *DARE* environment appears in Figure 1.

FIGURE 1: Sample *DARE* Screen



Indexed variations of the observed data set were used as input for the *DARE* environment by excluding selected groups of terms representing different term frequencies of occurrence. The data set with all terms and documents served as a benchmark from which comparisons were made. Subsets of the original set, which were input to the *DARE* system and tested, comprised: (1) all terms occurring more than one time (684 terms 3,911 postings); (2) all terms except the most frequently occurring term (1,640 terms, 4,748 postings), and; (3) all terms occurring less than 24 times (1,617 terms, 3,906 postings), corresponding to removal of roughly the same number of postings as in (1).

To normalize the results, which may be impacted by differences in the number of documents resulting from each combination, the authors employed the idea of a document space density (DSD) introduced by Salton (1989) and used by Korfhage (1997) as the basis for comparison. The DSD provides an indication of how densely documents are situated within the document space. It is defined as the mean of the sum of the similarity measures between each document and the document centroid:

where, n is the number of documents and $Sim(D_i, C)$ is the document D_i to centroid C similarity measure, and $Dist(D_i, C)$ (> 1.0) is the distance

$$DSD = \frac{\sum_{i=1}^{n} Sim(D_i, C)}{n}$$

where

$$Sim(D_i, C) = \begin{cases} \dfrac{1}{Dist(D_i, C)} & for\ Dist(D_i, C) \neq 0 \\ 1 & for\ Dist(D_i, C) = 0 \end{cases}$$

between document D_i and centroid C. The DSD can provide an important indicator of the characteristics of the document space. For the present study, the generated *DSD* values are used solely for comparison purposes. From a retrieval perspective, it is easier to distinguish documents from one another in a low-density space. The document space density involves more than just measurement issues, where the precision of the scale used allows one to still distinguish documents in close proximity to one another. In higher-density spaces, there is a greater likelihood for complete document overlap. In this situation, higher levels of scaling or more precise units of measurement cannot be used to distinguish one document from another.

To compare the influence of term distributions and indexing exhaustivity on document spaces, simulated data sets using different term frequency distributions and indexing exhaustivity levels based on the observed data set were tested. The observed term frequency and indexing exhaustivity distributions were first fitted to a simple Zipf and negative binomial models, respectively. Two variations of each distribution were then generated by altering parameter values so that the number of tokens generated for the term distribution and the number of terms generated for the indexing exhaustivity distribution represented average values above and below the observed means. The simulated sets consisted of three term distributions (shallow, observed, steep), representing an average of 4.0, 3.0, and 2.0 tokens per term type, and three indexing exhaustivity levels (low, observed, high) representing an average of 5.7, 9.7, and 13.7 terms assigned per document, respectively. The result was a 3 x 3 matrix, representing nine different combinations of indexing characteristics. The authors recognize the importance of term type co-occurrence dependencies, where specific terms may co-occur

with one another more frequently than chance would dictate. The authors chose to eliminate this confounding factor. With the different term frequency and indexing exhaustivity distributions used in each simulation combination, the incorporation of term dependencies for hypothetical data sets, for which there were no "real" equivalent data sets, could not be reasonably determined.

When constructing a set of hypothetical documents using computer simulation, an exit condition must be specified to end the simulation. Possible variables for this exit condition include the number of postings generated, the number of distinct term types, and the number of documents. The number of distinct terms was fixed for the investigation because both the indexing exhaustivity and the term distribution will impact the number of distinct terms, unlike postings, which are simply a by-product of term generation. Trial runs varying the number of term types generated did not result in notably different outcomes. Likewise, the number of documents may influence the number of distinct terms generated, but is not influenced by indexing exhaustivity. Also, any differences in the number of documents are accounted for by the *DSD* calculation.

The hypothetical document sets for each combination of term and exhaustivity distribution resulted in two data set variations: one that included all generated terms, and another that eliminated all singly occurring terms. Document space density values were then calculated using *DARE* outputs and were compared across the simulated data sets.

Results

Observed data set

Document space density values for the original data set and sets with the removal of different numbers of term types appear in Figure 2. The resulting density values reveal the greater influence of the removal of term types over term tokens on the organization of the document clusters. Removal of the most frequently occurring term, appearing in 25% of all documents, basically has no impact on the density of the document space.

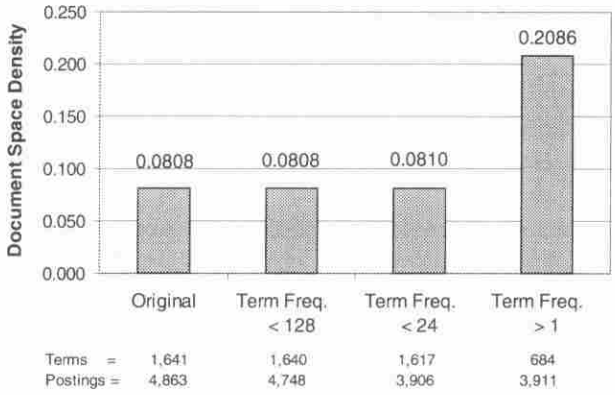


FIGURE 2: Document Space Densities for Different Term Subsets

The removal of a given number of term types occurring once has a much greater influence on the document-space density than the removal of an equivalent number of tokens for a smaller number of term types. In this case, those terms occurring 24 or more times, representing the 24 most frequently occurring term types were removed. Basically, the removal of an equivalent number of tokens for a much smaller number of types again produced very little change in the document-space density.

Simulated data sets

With a fixed number of index terms, the term and indexing exhaustivity distributions resulted in different numbers of documents and postings being generated for each simulation combination. For example, the difference in the number of documents generated between the low-exhaustivity/shallow-term distribution and the high-exhaustivity/steep-term distribution was a factor of more than four, but the numbers of term types used in each run were approximately equal.

Singly occurring terms were also removed from the simulated data sets and projected into the *DARE* environment. The resulting DSD values for each combination of indexing characteristics, for each set containing all terms and the removal of singly occurring terms appear in Figure 3. Note that the highest DSDs are observed in low-indexing exhaustivity environments with the removal of singly occurring terms. The term distribution also impacts the density, where steeper distributions (i.e., a

greater proportion of singly occurring terms) result in lower densities; however, with higher levels of exhaustivity, this influence decreases. Differences in DSD values across the indexing characteristics are still evident with the removal of singly occurring terms, although the values overlap across indexing exhaustivities.

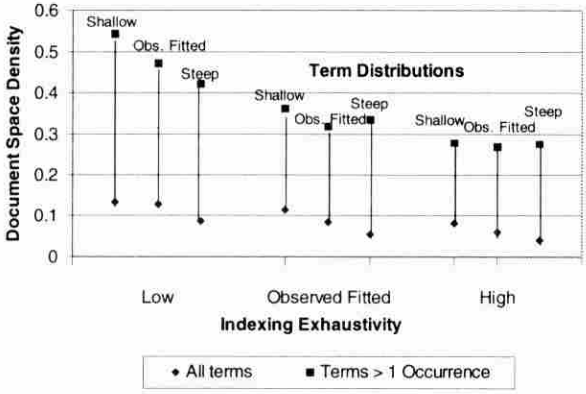


FIGURE 3: Document Space Densities for Simulated Data Sets

Figure 3 demonstrates the differences in DSD values across indexing combinations, but it does not provide a good sense of the change in values between the complete term set and the removal of singly occurring terms. Figure 4 outlines the changes in DSD values with the removal of singly occurring terms. The biggest changes occur with the lowest exhaustivity, where, with fewer terms assigned per document, the individual influence of assigned terms is greater. The term distribution also plays a role, but the impact varies based on the exhaustivity. The shallowest distribution results in the largest change across different exhaustivity levels, while the steepest distribution results in the smallest change.

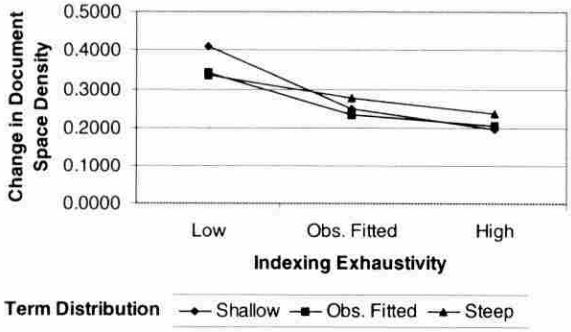


FIGURE 4: Change in Document Space Density for Simulated Sets

Discussion

From the system representation and organization perspective, the presence of singly occurring terms is important in defining the density of the document space, and more specifically in distinguishing individual documents. But from the user perspective, these terms may or may not be important to the user's retrieval needs. They do not cast a broad net for querying purposes, but are able to pinpoint specific documents. Based on evidence collected from an on-line public access catalogue (Wolfram 1992), when cumulated, singly occurring terms were used more in queries than the most frequently occurring terms, although at the individual term level, their use was smaller than for specific high-frequency terms. It is unknown whether this term selection behaviour is observed for other IR environments, where the user characteristics, information needs, retrieval expectations, and content indexed varies. The importance of singly occurring terms to the user requires further investigation; however, the observation lends some evidence to the importance of the use of singly occurring terms from the user's perspective as well. High-frequency terms may be viewed as more useful for the retrieval of larger sets, but they play only a small role in defining the document-space characteristics, at least when these terms are evenly distributed across the document set. In the case of the observed data set, the impact of the removal of term types was more important than the removal of an equivalent number of term tokens. In systems where real-time indexing of document sets takes place, the elimination of singly occurring terms helps to reduce the computational burden associated with document-space calculations in vector space model implementations. However, in eliminating these terms, the designer will be removing unique characteristics that define the document set. So, for automatic indexing purposes, it is better to keep the singly occurring terms. To maximize document distinctiveness to aid in retrieval, the system designer would be better off deleting high-frequency terms instead of low-frequency terms, if indexing policy dictated the removal of any term types.

The proportion of singly occurring terms characterized by the distribution of term frequencies is not the only factor that impacts the characteristic of the document space. Different levels of indexing exhaustivity play a stronger role in ultimately determining the document-space density. Higher levels of exhaustivity result in less-dense environments, indicating greater distinctiveness between documents, even with diffe-

rent term distribution characteristics. Also, the impact of the removal of singly occurring terms becomes less significant with higher exhaustivities. So, to maintain a lower document space density in the absence of singly occurring terms, higher exhaustivity levels for remaining terms can be implemented during system indexing. However, given the large impact of the singly occurring terms on the document-space density, a higher exhaustivity can reduce, but not eliminate this impact in the absence of the singly occurring terms.

The authors are cautious in drawing sweeping conclusions regarding indexing in general due to the limitations of the study. The size of the observed and simulated data sets is relatively small in comparison with commercial systems. Again, the intent here is not to model the entirety of a large document set but to model a subset from which more exhaustive searching can take place. The authors' intent was also to demonstrate a methodology for studying document-space characteristics in a visual environment using simulation. Simulation studies provide the advantage of exploring potential outcomes under different circumstances at a fraction of the cost associated with collecting numerous data sets from actual systems. If notable outcomes are observed, they may be confirmed with actual data sets.

Conclusion

Visualization techniques have made it possible to study established IR approaches in new ways. Through systems that permit the visual display of document sets, it becomes possible to test what may have been long suspected from intuition but was difficult to demonstrate. The present research has demonstrated the influence of different indexing characteristics on the document space of a small set of documents. The many singly occurring terms indexed within the test document set and simulated sets play a key role in defining the characteristics of the document space, although they may or may not be useful from the user's perspective for searching. Through the use of index modelling and computer simulation techniques, the authors were also able to demonstrate the impact of different term distributions and indexing exhaustivity levels on the resulting document space. Although both the term distribution and exhaustivity impact the document space density, the exhaustivity is more influential in defining the density.

Future research should examine the influence of intra-document term weights assigned to term tokens. Terms weights assigned to each occurrence of a term across documents were treated as equal in the present study since no data were available on the weight of each term within a given document. Even though a term type may occur within many documents, the concentration of occurrences within a given document may result in a higher weight being assigned to the term in one document when compared with another where the term is less prominent. The identification of significant terms based on their concentration within documents can then be used to decide which terms of equal frequency of occurrence within a document set, with different concentrations across documents, should then be included or excluded during indexing.

References

Benford, S., C. Greenhalgh, D. Snowdon, R. Ingram, and I. Knox. 1995. VR-VIBE: A virtual environment for co-operative information retrieval. In *Eurographics* 95: 123–34.

Bird, P. R. 1974. The distribution of indexing depth in documentation systems. *Journal of Documentation* 30, no. 4: 381–90.

Burnett, J. E., D. Cooper, M.F. Lynch, P. Willett, and M. Wycherley. 1979. Document retrieval experiments using indexing vocabularies of varying size. I. Variety generation symbols assigned to the fronts of index terms. *Journal of Documentation* 35, no. 3: 197–206.

Fedorowicz, J. 1982. A Zipfian model of an automatic bibliographic system: An application to Medline. *Journal of the American Society for Information Science* 33: 223–32.

Fekete, J. D., and N. Dufournaud. 2000. Compus: Visualization and analysis of structured documents for understanding social life in the sixteenth century. In *Proceedings of the fifth ACM conference on digital libraries,* 47–55. San Antonio: ACM.

Hearst, M. A. 1995. TileBars: Visualization of term distribution information in full text information access. In *Proceedings of CHI'95 human factors in computing systems,* 59–66. Denver, CO: ACM.

Helfman, J. I. 1994. Similarity patterns in language. In *Proceedings 1994 IEEE symposium on visual languages*, 173–75. St. Louis, MO: IEEE.

Kim, H. and R. Korfhage. 1994. BIRD: Browsing interface for the retrieval of documents. In *Proceedings 1994 IEEE symposium on visual languages*, 176–77. St. Louis, MO: IEEE.

Korfhage, R. 1997. *Information storage and retrieval.* New York: Wiley Computer Publications.

Liu, Y. H., P. Dantzig, M. Sachs, J.T. Corey, M.T. Hinnebusch, M. Damashek, and J. Cohen. 2000. Visualizing document classification: A search aid for the digital library. *Journal of the American Society for Information Science* 51, no. 3: 216–27.

Nelson, M. J. 1989. Stochastic models for the distribution of index terms. *Journal of Documentation* 45, no. 3: 227–37.

Nelson, M. J., and J.M. Tague. 1985. Split size-rank models for the distribution of index

terms. *Journal of the American Society for Information Science* 36: 283–96.

Nuchprayoon, A., and R.R. Korfhage. 1994. GUIDO: A visual tool for retrieving documents. In *Proceedings 1994 IEEE computer society workshop on visual languages*, 64–71. St. Louis, MO: IEEE.

Olsen, K. A., and R.R. Korfhage. 1994. Desktop visualization. In *Proceedings 1994 IEEE symposium on visual languages*, 239–44. St. Louis, MO: IEEE.

Salton, G. 1989. *Automatic text processing: The transformation, analysis, and retrieval of information by computer.* Reading, MA: Addison-Wesley.

Shneiderman, B., D. Feldman, A. Rose, and X.F. Grau. 2000. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the fifth ACM conference on digital libraries*, 57–62. San Antonio: ACM.

Small, H. 1999. Visualizing science by citation mapping. *Journal of the American Society for Information Science* 50, no. 9: 799–813.

Tague, J., and P. Nicholls. 1987. The maximal value of a Zipf size variable: Sampling properties and relation to other parameters. *Information Processing & Management* 23, no. 3, 155–70.

Willett, P. 1979. Document retrieval experiments using indexing vocabularies of varying size. II. Hashing, truncation, digram and trigram encoding of index terms. *Journal of Documentation* 35, no. 4: 296–305.

Wolfram, D. 1992. Applying informetric characteristics of databases to IR system file design, Part I: informetric models. *Information Processing and Management* 28, no. 1: 121–33.

Wolfram, D., and J. Zhang. 2002. An investigation of the influence of indexing exhaustivity and term distributions on a document space. *Journal of the American Society for Information Science and Technology 53, no. 11: 943–952.*

Young, D., and B. Shneiderman. 1993. A graphical filter/flow representation of Boolean queries: A prototype implementation and evaluation. *Journal of the American Society for Information Science* 44, no. 6: 327–39.

Zhang, J. 2000. A visual information retrieval tool. In *Proceedings of the 63rd annual meeting of the American Society for Information Science*, 248–57. Medford, NJ: Information Today, Inc.

Zhang, J. 2001. TOFIR: A tool of facilitating information retrieval. Introducing a visual retrieval model. *Information Processing and Management* 37, no. 4: 639–57.

Zhang, J., and R. Korfhage. 1999. *DARE:* Distance and angle retrieval environment: A tale of the two measures. *Journal of the American Society for Information Science* 50, no. 9: 779–787.

Zhang, J., and D. Wolfram. 2001. Visualization of term discrimination analysis. *Journal of the American Society for Information Science and Technology* 52, no. 8: 615–27.