An Investigation of the Influence of Indexing Exhaustivity and Term Distributions on a Document Space

Dietmar Wolfram and Jin Zhang
School of Information Studies, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201.
E-mail: dwolfram@uwm.edu, jzhang@uwm.edu

The authors investigate the influence of index term distributions, and indexing exhaustivity levels on the document space within a visual information retrieval environment called DARE. Using combinations of three levels of term distributions (shallow, observed, steep) and indexing exhaustivity (low, observed, high), hypothetical document sets were generated and projected onto the DARE environment. The results from the simulated document sets demonstrate the importance of term distribution and exhaustivity characteristics on the density of document spaces and their implications for retrieval, particularly when different term weighting schemes are used. The results also demonstrate how different combinations of exhaustivity and term distributions may result in similar document space density characteristics.

Introduction

The effectiveness of an information retrieval (IR) system is largely determined by its ability to discriminate between the documents it indexes, so that documents deemed most relevant to a user's query can be distinguished from those that are less relevant. By incorporating different indexing and retrieval techniques, IR systems are able to maximize identification of potentially relevant documents. Indexing rules, such as truncation and term weighting will dictate which contents of a document are included for retrieval. These rules will also ultimately shape the observed regularities that result from indexing, namely how frequently specific terms are included in the index, as well as how many index terms are used to identify a given document (indexing exhaustivity). Occurrences in term assignment may be summarized in the form of term frequency and indexing exhaustivity distributions. Although useful for summarizing the nature of indexes within IR systems, these distributions do not reveal the influence of term assignment on the relationships among individual documents.

Recent developments in IR system design and document representation have made it possible to represent the interrelationships among documents using information visualization techniques. Visualization affords the opportunity to summarize complex relationships in a comprehensible manner for the system user, something that is not possible when presented with a single value or linear lists of ranked documents from an IR system. For years, researchers have been studying the impact of different indexing methods on retrieval effectiveness, and have developed single outcome measures of system entropy. Single values, although useful as summarizations for comparison purposes, do not shed light on the range of values encountered.

The purpose of the present study is to explore the influence of indexing characteristics (indexing exhaustivity, term distributions, and term weight assignment methods) on the document space in a vector-based IR environment. These characteristics define the relationships among indexed documents and, ultimately, will influence the retrieval process. Two primary research questions guide the present research. First, how do different indexing exhaustivity and index term distributions resulting from the system indexing process impact document cluster organization in the document space? Second, how are document clusters impacted in these different indexing environments when different methods of term weights are assigned to index terms?

Related Research

The present study incorporates several areas of information science research. The authors use information visualization to represent the complex relationships that exist among documents within an IR system and rely on experimentation with different system characteristics. Underlying the studied system are indexing assumptions that rely on empirical regularities based on term assignment and their distribution studied in informetrics. Finally, the study em-

Received January 3, 2002; revised March 22, 2002; accepted March 22, 2002

© 2002 Wiley Periodicals, Inc.

ploys computer simulation methods to test hypothetical environments incorporating different indexing strategies.

Information visualization uses interactive visual representations of abstract data to amplify cognition, analyze data, and process data within a visual environment. It simplifies information representation in information systems by applying visual processing to abstract information. Information visualization combines aspects of scientific visualization, 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, Greenhalgh, Snowdon, Ingram, & Knox, 1995; Kim & Korfhage, 1994; Olsen & Korfhage, 1994), or visually demonstrate attributes of documents or semantic relationships (Fekete & Dufournaud, 2000; Helfman, 1994; Shneiderman, Feldman, Rose, & Grau, 2000), or visualize internal mechanisms for retrieval processing (Nuchprayoon & Korfhage, 1994; Young & Shneiderman, 1993; Zhang, 2001; Zhang & Korfhage, 1999). However, the application of visualization techniques in the information sciences, in addition to its applications in information seeking, may be applied to information analysis. For instance, visualization can be applied to term discriminative capacity analysis (Zhang & Wolfram, 2001), document classification analysis (Liu et al., 2000), citation analysis (Small, 1999), and full-text analysis (Fekete & Dufournau, 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, Cooper, Lynch, Willett, and Wycherley (1979), examined the effect of the size of controlled index term vocabularies on retrieval using the Cranfield test collections. The authors 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 aspects 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 number of terms occur with great frequency and a large number of terms occur only once or twice. Researchers have attempted to model the observed term distribution behavior by fitting observed data sets to theoretical distributions. Houston and Wall (1964) applied a lognormal model to a set of term frequency data with little success. Bennett (1975), using a size-frequency data set, modeled index term data from two bibliographic databases using a Zipf distribution. The author concluded that the poor fits in the tail were due to the relatively small size of the index term set, but subsequent research by others would show this was often the case even with larger numbers of terms. Generalizations of the Zipf distribution, such as a three-parameter Mandelbrot-Zipf have been shown to provide better fits than the traditional Zipf (Wolfram, 1992a). 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.

Another important feature of IR systems important in modeling 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 (see later). He also attempted to fit the continuous lognormal distribution with limited success. Since then, researchers such as Nelson and Tague (1985) and Wolfram (1992a) 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. The steepness of ascent and descent of the theoretical distributions did not always adequately model those of the observed distributions. However, for general modeling purposes such as computer simulations of IR systems where means and variances are known, the negative binomial distribution is usually the best to use.

With appropriate models of IR system content, one may develop computer simulations to test system performance under different circumstances. Simulation serves as a useful tool to test hypothetical models of system processes without the expense and resources of having to develop different systems. Over the past 30 years, a number of researchers have developed simulation models to represent the processes within IR systems. Cooper (1973) developed one of the earliest comprehensive models. The model consists of five parts; thesaurus generator, document generator, query generator, search routines, and evaluation routines. Due to the complexities inherent in the interactions of different IR system components, which could not be adequately modeled, the author concluded that the simulation model required additional components to be an effective evaluative tool. Tague, Nelson, and Wu (1981) examined key difficulties with the overall simulation of retrieval systems. Their simulation model could generate both a document description based on the term distribution, indexing exhaustivity, and cooccurrence of terms, and also generated a query set. More recently, Wolfram (1992b) developed a simulation model to test optimal file structures for inverted file information retrieval systems. The model included data for terms used per query, index term distributions, and distributions

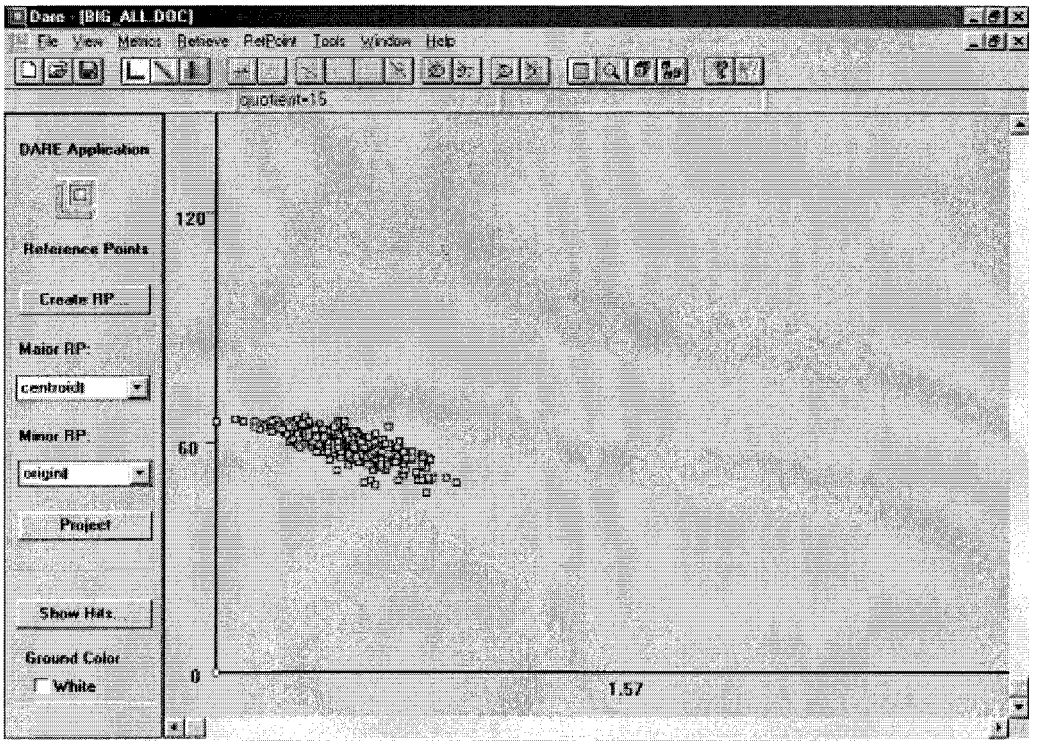

FIG. 1. Sample DARE screen shot.

for index term selection. By varying the parameter values index term distributions and distributions for term selection, the author was able to test the performance of various file structures under different database and search characteristic environments.

The present study builds on previous IR visualization and simulation research by examining the influence of different term and indexing exhaustivity distributions on the document space of a visually based IR environment.

Method

Descriptor frequency and indexing exhaustivity data for 500 bibliographic records dealing with library and information science were used from the NTIS database. The data source was selected for its availability, having been used in previous research by one of the investigators, its use of a controlled vocabulary, and level of descriptor assignment. The observed term and term exhaustivity distributions were plotted, with means, and maximal values calculated for each. The characteristics of this data set served as the basis for the development of hypothetical models incorporating term distribution and exhaustivity features. Although the data set used is small by today's large gigabyte database standards, the nature of the environment and the processing required to project the document set onto a visual space required the use of a manageable number of index terms and documents for processing in the visual DARE system.

Variations of the observed data were used as input for the DARE visual information retrieval environment. The distance-angle-based visual tool *DARE (Distance Angle Retrieval Environment)* (Zhang, 2000; Zhang & Korfhage, 1999), is a two-dimensional visual retrieval tool, consisting of a graphical representation of the visual distance and the visual angle as the *X*-axis and *Y*-axis respectively. A screenshot of a sample DARE output appears in Figure 1.

Given two reference points in a vector-based document space (one is the major reference point that is more important to user's information need and another is the minor reference point that is less important to the information need), these two reference points can determine a line in the document vector space. It is clear that any indexed document can be located in the document space. Observe that a document in the document vector space corresponds to the two important parameters vis-à-vis the defined reference points. One is the visual distance, defined as the distance from that document to the major reference. The other is the visual angle, defined as the angle formed by the document and the minor reference point against the major reference point (or by the two lines determined by the document and the major reference point as well as the minor reference point and the major reference point, 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 whose *X*-axis and *Y*-axis are

defined as the visual angle and visual distance respectively. The *X*-axis of the visual area ranges from zero to π due to the symmetrical characteristic against the reference line determined by the major and minor reference points in measuring a visual angle of a document. Because there is no limitation on a visual distance of a document, the *Y*-axis can range from zero to infinity theoretically.

The document vector space *V* is defined in Equation 1.

$$V = \begin{Bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{Bmatrix} \quad (1)$$

The columns of the vector *V* are indexing terms in a document collection. The rows of the vector are documents that are indexed by the indexing terms. The number of different indexing terms in the collection is *n* and the number of documents is *m*. t_{ij} is defined as weight of indexing term *j* for document *i*.

The document centroid is defined as the average document situated in the center of the document space. In this case, the document centroid can be expressed as:

$$C = (c_1, c_2, \ldots, c_n) \quad (2)$$

where

$$c_i = \frac{\sum_{j=1}^{m} t_{ji}}{k}, \quad (i = 1, \ldots, n), \quad (3)$$

Here, *n* is the number of terms in the collection and *k* is the number of nonzero elements in $\sum_{j=1}^{m} t_{ji}$.

To effectively observe the change in the distribution of document clusters in the visual space, the centroid and origin of the document space are assigned as the major reference point and minor reference point, respectively. In this case, observers can measure the average distance between documents and the centroid within the DARE visual environment after all documents are projected onto the space. The average distance is the distance from the origin of the visual space to the projected document body. The collective average distance is then converted to the similarity value. This strategy suggests that observers can observe whole document clusters from the center of document space because the centroid is selected as the major reference point. Because our study focuses on the impact of distance rather the direction, the selection of the minor reference point is not important.

The investigators examined the impact of different term distributions and exhaustivity levels on the distances generated in the document space of DARE. This may be accomplished with different combinations of term frequency and indexing exhaustivity distributions. Nine simulation models were developed around the observed data sets, representing a 3 × 3 matrix of term distributions (shallow, observed, steep) and indexing exhaustivity levels (low, observed, high). The term distribution of the observed data set was assumed to follow a simple Zipfian model and served as a template for model development. The simple Zipf distribution provides a parsimonious model with parameters that play distinct roles in influencing the shape of the distribution. The size-frequency form of the simple Zipf distribution was used to model the data (Tague & Nicholls, 1987):

$$f(x) = \frac{a}{x^b}, \quad x = 1, 2, \ldots, x_{\max} \quad (4)$$

where *f(x)* represents the proportion of terms of size *x*, while *a*, *b*, and $x_{\max}$ represent constants. The *a* parameter defines the shape of the low end of the distribution, dictating the proportion of values; while the *b* parameter determines how steeply the distribution descends. The $x_{\max}$ value was set to equal the largest occurrence type within the observed data set. Parameters *a* and *b* were estimated for the observed data set for each model using combinations of the observed term types so that the mean frequency of terms equaled the observed mean. The parameters for the two hypothetical models of term distributions were calculated using iterative methods that relied on the goal average frequency for that model so that the density function of the term distribution totaled 1.0 when cumulated to $x_{\max}$. At the tail end of the distribution, where gaps in the observed term types existed, probabilities for a given term type were cumulated between types. A cumulative form of the density function was tabulated for each distribution based on the number of postings (i.e., tokens) associated with each term type instead of the frequency of each type. Using type frequency alone would underrepresent the selection of terms occurring more than once.

Indexing exhaustivity distributions are usually unimodal with modes greater than one term per document (Bennett, 1975; Nelson & Tague, 1985; Wolfram, 1992a). To model indexing exhaustivity, a negative binomial distribution was assumed:

$$p(x) = \binom{x + v - 1}{v - 1} p^x (1 - p)^v \quad (5)$$

where *v* and *p* represent parameters. Each parameter may be estimated using moment estimators based on the observed distribution mean and standard deviation. Again, goal means for term exhaustivity were set for each model, with the observed maximum number of terms assigned per document being used as the model maximums. Because the negative binomial distributions starts at 0 and the minimum number of terms assigned per document was 1, the number of terms generated was shifted up by 1. A zero-truncated form of theoretical distributions could also have been used to generate theoretical values. Given the low likelihood of

zero values for this distribution when the mode is large, the differences between the two approaches is small, when compared to hyperbolic distributions with modes at the lowest x-value. A cumulative probability distribution was then generated for each indexing exhaustivity model based on the model probability function.

A simulation program was written for the Microsoft Access™ database management system environment. The program generated tables of hypothetical documents and terms based on the term distribution and exhaustivity models. For each simulation run, exhaustivity values were first generated for each document, with term sizes generated for each document term. The authors acknowledge the influence of term dependence on term cooccurrence. However, the models used assume independence of term occurrence because the varying impact of dependence that results from the different term distributions and indexing exhaustivity levels was unknown; also, the vocabulary being generated was hypothetical. The program continued to generate additional documents until the number of distinct terms generated approached the observed number of terms. Term totals may not equal the observed number as a result of the number of terms generated by the last document. Once each run was finished, database table contents were converted to file formats compatible with the DARE system. The document spaces created by the hypothetical document sets were visually inspected and compared for changes in distances of document clusters from the document set centroid.

In building a hypothetical set of indexed documents, an exit condition must be specified to end the simulation. Candidate variables for this include the number of postings generated, the number of different terms, and the number of documents. The number of index terms was fixed for this study because both the indexing exhaustivity and the term distribution will impact the number of distinct terms, whereas postings are simply a by-product of term generation, Likewise, the number of documents may influence the number of distinct terms generated but is not influenced by indexing exhaustivity.

Three sets of simulation runs were performed incorporating classic term weighting schemes. The first simulation model used inverse document frequency (*idf*) term weight, calculated as:

$$idf_k = \log\left(\frac{N}{freq_k}\right) \tag{6}$$

where freq_k is the number of documents in which term k appears and N is the total number of documents.

The second set of simulation runs incorporated intradocument term weights only based on term frequency (*tf*):

$$tf_{ik} = \text{freq}_{ik} \tag{7}$$

where freq_{ik} is the number of times in which term k appears in document i. Because the original term set used consisted of descriptors with no intradocument frequency data, term weights were calculated based on an assumed Zipfian distribution of term frequency of occurrence with the document. Although the term frequencies within the document were based on an assumed Zipfian distribution, the generation of the different terms within and across documents was based on the three simulated distributions.

The third model, which was of primary interest, used the inverse document frequency term weights based on both inter and intradocument term frequencies (*tf–idf*):

$$tf - idf_{ik} = \text{freq}_{ik} * \log\left(\frac{N}{\text{freq}_k}\right) \tag{8}$$

Interdocument frequencies were based on the distributions in the *idf* model. Intradocument term weights were calculated using the similar Zipfian assumption as in the *tf* model. The *idf* and *tf* term weighting schemes were used to identify overall impact on the document clusters. More detailed analysis of the *tf–idf* scheme was undertaken, with five runs for each combination being conducted so that average outcomes could be tabulated.

For each simulation result, a screenshot of the resulting document set was taken. Document, term, and posting totals for each run were tabulated. Average document to centroid distances, providing an indication of document cluster size, were determined from the screenshots. Because document distances were impacted by the number of documents generated in each run, a normalized method to compare the impact of each term factor on the document space was needed. The authors employed the idea of a document space density introduced by Salton (1989) and used by Korfhage (1997) to remove the influence from the size of the document sets. The document space density (*DSD*) provides an indication of how densely documents are organized within the document space. It is defined as the mean of the sum of the similarity measures between each document and the document centroid:

$$DSD = \frac{\sum_{i=1}^{n} \text{Sim}(D_i,C)}{n} \tag{9}$$

where, n is the number of documents and $\text{Sim}(D_i, C)$ is the document D_i to centroid C similarity measure such that

$$\text{Sim}(D_i, C) = \begin{cases} \dfrac{1}{\text{Dist}(D_i, C)} & \text{for Dist}(D_i, C) \neq 0 \\ 1 & \text{for Dist}(D_i, C) = 0 \end{cases} \tag{10}$$

where and $\text{Dist}(D_i, C)$ (>1) is the distance between document D_i and centroid C. The generated *DSD* values were used solely for comparison purposes within each set of simulation runs.

TABLE 1. Zipf parameter estimates for term distributions.

Term distribution characteristic	*a*	*b*	Goal for mean term frequency (x_{max} = 128)
Shallow	0.565	1.865	4.0
Observed (Fitted)	0.630	2.053	3.0
Steep	0.716	2.374	2.0

TABLE 2. Negative binomial parameters for indexing exhaustivity distributions.

Exhaustivity characteristic	*p*	*v*	Goal for term mean frequency (maximum = 26)
Low	0.458	4	5.7
Observed (Fitted)	0.309	4	9.7
High	0.7995	51	13.7

Results

The observed term distribution contained 1,641 terms with 4,863 postings. The mean term frequency for the observed data set was 3.0 tokens per term. Goal mean frequencies for the shallower and steeper term distributions were set so that the mean tokens per term were 1.0 above and below the observed mean, respectively, resulting in shallower and steeper term distributions, respectively. Hypothetical term distributions were modeled based on these assumptions. Resulting parameter values for the Zipf model appear in Table 1. The shape of the low end of each distribution appears in Figure 2.

The same approach was used for the indexing exhaustivity data using the negative binomial distribution. The observed distribution mean of 9.7 terms per document was used as the basis of the observed (fitted) distribution. Moment estimators for fitting the model to the data were calculated and then adjusted using iterative combinations of parameter values, to minimize chi-square goodness-of-fit values. Goal means for the low and high exhaustivity values were set at values representing four below and four above the observed means respectively. The investigators felt that going much lower than an average of five terms assigned per document would not provide much distinction between documents generated. The maximal value for each model was fixed at the observed value of 26 terms per document. Final outcomes for the indexing exhaustivity models used appear in Table 2.

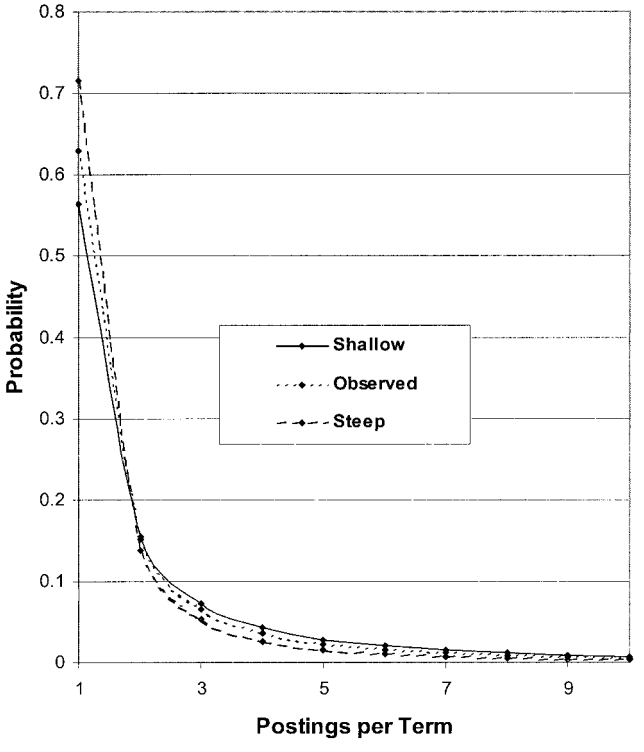

FIG. 2. Resulting term distributions.

Characteristics of the document sets varied with each simulation combination. The generated number of distinct terms was always at or within several terms of the 1,641 terms found in the observed set; however, the numbers of documents and postings varied as a result of the term distribution and exhaustivity levels. For example, the number of documents generated in the simulation runs with low exhaustivity and shallow term distributions resulted in document sets almost five times as large as those runs with high exhaustivity and steep term distributions. In the former, the additional documents were needed to generate a sufficient number of distinct terms. However, with the latter, the steep term distribution and higher exhaustivity permitted the observed number of terms to be generated using fewer documents. For the same combinations, the total number of postings generated differed only by a factor of two. The validity of the document set characteristics for each simulation combination was verified by comparing the average number of terms per document and tokens per term type generated to the expected number for each model. In each case, the difference between the simulated and expected values was less than 5%, indicating that the generated systems did indeed represent the characteristics of the model selected.

Sample output from the DARE system appears in Figure 3 for three model combinations. Because DARE can display both distance and angle measures, only data point values on the y-axis, which are distance dependent, are meaningful. Variations in the x-axis, dealing with angle measures are not meaningful in this context. The clustering behavior of the documents was relatively uniform across all simulation combinations, with little variation in the individual document-to-centroid distances within a given run, but with more notable variations in the average document-to-centroid distances across simulation combinations.

Based on the simulation runs, the changes in the document space density patterns were similar for each of the model combinations, although the actual density values differed widely for each term weighting scheme. In each

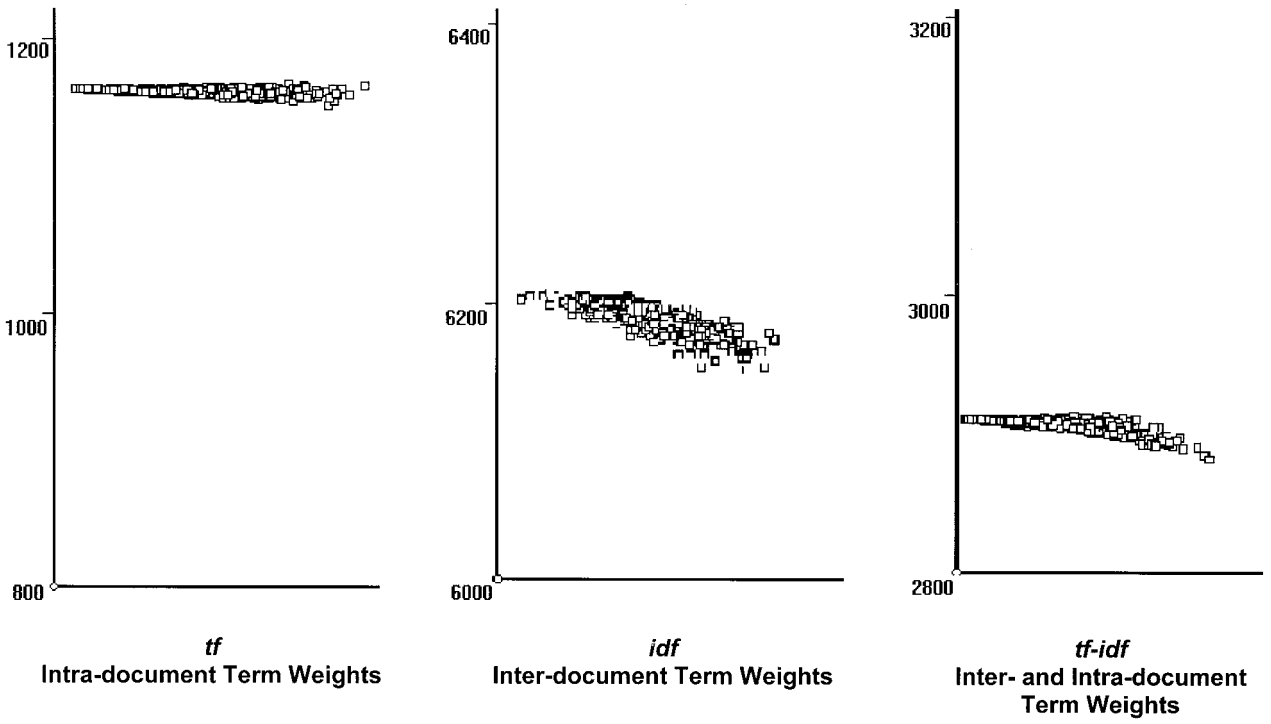

FIG. 3. Sample DARE outcomes for each model using fitted forms of the observed term and indexing exhaustivity distributions.

case, regardless of the term weighting scheme, steep term distributions resulted in the lowest document space densities, with the shallow term distribution producing the highest densities (i.e., the most tightly clustered document sets), confirming that fewer tokens per term type result in a more distinctive document set. Similarly, indexing exhaustivity also contributed to the density values, so that higher exhaustivity levels resulted in lower density spaces.

The interdocument inverse document frequency term weight (*idf*) system runs resulted in the most diffuse document spaces of the three term weighting schemes. The roles of both indexing exhaustivity and term distribution in distinguishing documents from one another is evident (Fig. 4).

For the intradocument term frequency term weight model (*tf*), additional model assumptions had to be made. Because no intradocument term weight data were available from the initial data set, term weights were modeled based on a Zipf frequency distribution with parameters $a = 0.646$ and $b = 2.1$, with a maximal value of 50. Admittedly, this is an arbitrary assignment, but the investigators did not want the maximal value to be too large or that the term distribution be too steep, resulting in few frequently occurring terms. The distribution was truncated so that only term frequencies of at least five times were used, based on the assumption that a higher candidate term frequency within the document would indicate the term's worthiness of inclusion as an index term. The same pattern of influence by the interdocument term distribution and indexing exhaustivity is observe as for the *idf* model, although the density values are much higher (Fig. 5).

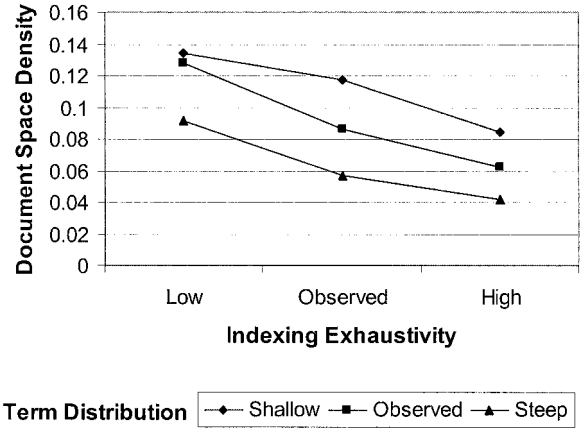

FIG. 4. Document space density comparison *idf* model.

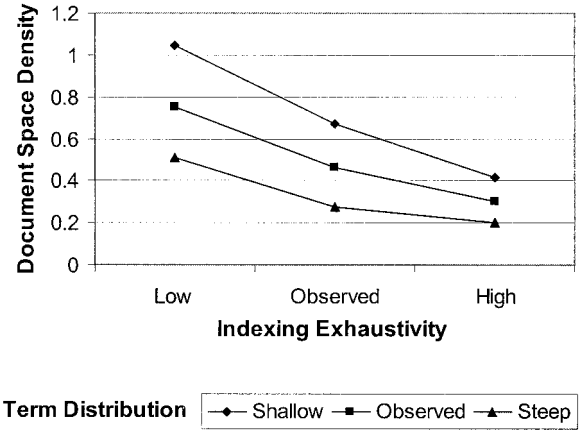

FIG. 5. Document space density comparison *tf* model.

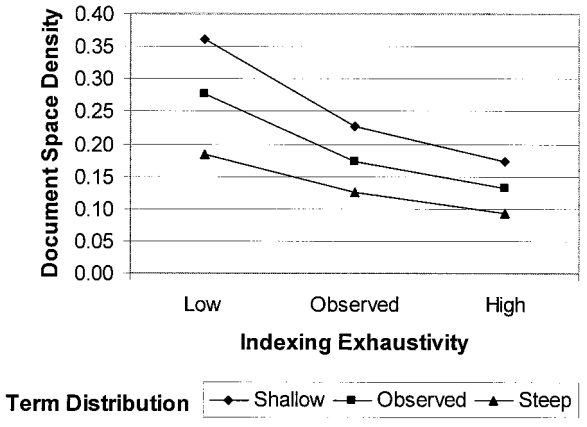

FIG. 6. Document space density comparison *tf–idf* model.

Results of the *tf–idf* simulation model that integrated *idf* and *tf* document term weights appear in Figure 6. The same intradocument term frequency distribution was used as in the *tf* term weight model. Sample DARE output for three simulation combinations appear in Figure 7. Density values were intermediate between those of the *idf* and *tf* term weighting schemes.

The small, but observable, variability in document-to-centroid distances within the document clusters for each model raises questions whether the indexing characteristics differed between those that were situated closer to the centroid and those that were farther away. The DARE system permits the selection of documents within the cluster so that the identity of the documents may be revealed. The authors isolated small subsets of 15 to 30 documents at the highest and lowest distances for selected simulation combinations for each model. *t*-Tests were performed on the average term size and indexing exhaustivity values for the documents at each extreme. No significant differences were found in the characteristics of each document set, indicating that term frequencies within documents and indexing exhaustivity levels alone did not impact where documents were located in the cluster.

Discussion

Visualization provides advantages for studying IR document spaces that are not inherent in other analysis methods. First, through visualization, the special changes of document distribution can be visually displayed. Second, it allows one to readily compare the characteristics of document clusters using simple visual inspection, something that is not possible when using a single summary value of a document space without a visual cue.

For a document space, the indexing exhaustivity and the term distribution will impact retrieval by defining the document space. From an information retrieval perspective, when the document space becomes denser, that is, the average number of documents per spatial unit becomes larger, the ability for a retrieval system to differentiate one document from other documents in the document space decreases. For instance, the distance retrieval model determines a hypersphere in the document space where the size of the sphere is based on the radius, which is defined as the threshold. When the document space becomes denser, more documents may be included or excluded from the contour after the threshold of retrieval increases or decreases a standard measure unit. In this sense, retrieval results would be more sensitive to the change in threshold of an information retrieval model.

Low exhaustivity and shallow term distributions produce fewer distinctions among documents because fewer terms are assigned per document and more common terms are shared among documents, resulting in more terms with a low term weight. Higher exhaustivity provides additional opportunities for extra, distinctive terms to be added to a document. Similarly, a steeper term distribution, where a lower average rate of specific term assignment to the doc-

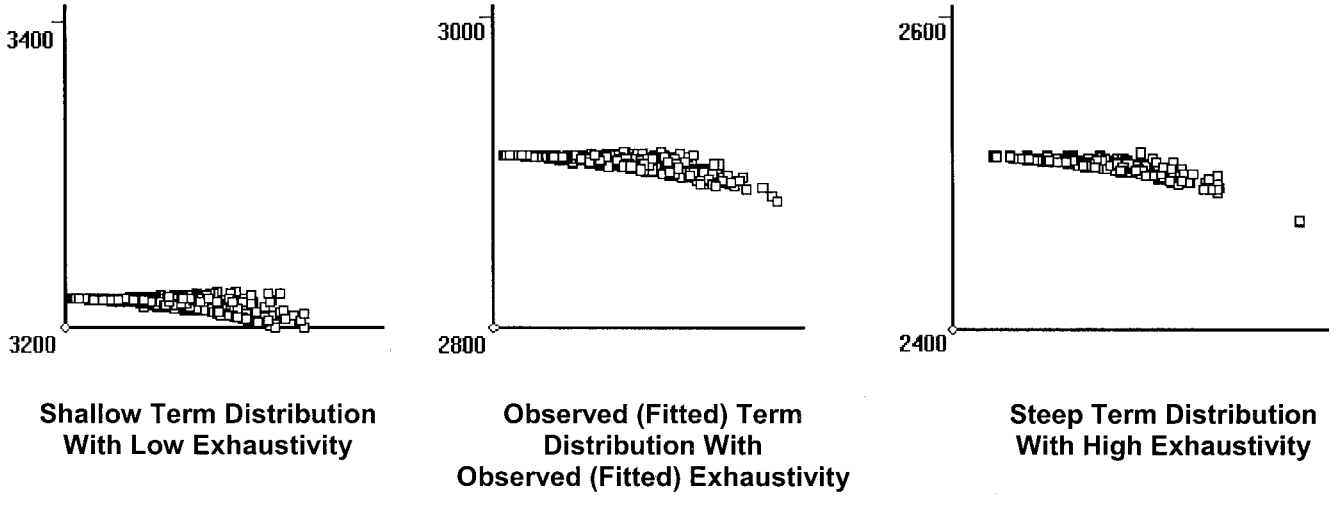

FIG. 7. DARE output for three simulation model combinations *tf–idf* model.

ument set is found, adds to the document's distinctiveness. With high exhaustivity to more completely describe a document and steeper term distributions, defining more unique sets of terms, the lowest document densities are encountered, making it easier to distinguish documents from one another. The results for each model demonstrate that similar document space densities may be achieved with different combinations of indexing exhaustivity and term distributions. For example, a high exhaustivity/shallow term distribution combination and a low exhaustivity/steep term distribution combination resulted in similar document space density levels.

The influence of each term weighting scheme is revealed in the different levels of document space density. The *idf* term weighting scheme resulted in the lowest densities. This is somewhat surprising, because the added distinctiveness of term frequencies within documents is not taken into account. However, because the classic *tf–idf* formula relies on raw term frequencies that have not been normalized, the influence of the term frequency within the document on the overall term weight is very strong. To retain lower document space densities in the *tf–idf* term weighting environment, the impact of the *tf* portion should be softened, for example by normalizing the raw frequencies, or the impact of the *idf* portion of the formula should be strengthened by reducing the base of the logarithm.

As with any experimental study there will be limitations resulting from constraints placed on the data to control for confounding factors. The present study relied on a fairly small set of documents containing a fixed number of index terms. A manageable number of terms were needed from a computational perspective to be compatible with the DARE environment. The assumption of indexing independence, which will impact the assignment of terms to documents, has been long known not to hold true, but in the absence of empirical evidence to support the levels of dependence for different exhaustivity and term distribution environments, this can only be guessed. Therefore, it was removed as a confounding factor. The inability to fix all system characteristics (terms, postings, documents) by changing indexing characteristics, raises the question whether document space characteristics were impacted by the variations in postings and or documents, and not necessarily the number of terms. However, by normalizing the results through the use of document space densities, as opposed to strict distance values, the influence of these differences, particularly in the number of documents is reduced.

The findings have implications for retrieval system analysis and design. Indexing exhaustivity should be maximized when indexing documents. Although this is intuitive, the simulations demonstrated how higher levels of exhaustivity impact the document space density positively, even in term distribution environments with shallower term distributions. To maintain document distinctiveness, the number of index terms should continue to grow. From a document representation perspective, this may allow greater distinctiveness between the documents. From a user search perspective, this will increase precision at the expense of recall. But, with today's IR systems indexing many millions of documents, added recall may not be what searchers are looking for.

Conclusions

Indexing methods will not only dictate how documents may be retrieved, but will also define how documents relate to one another. The present study has explored the impact of term assignment on a document space using different combinations of term frequency distributions and indexing exhaustivity levels. The importance of term weighting methods in defining the document space in conjunction with term assignment is also significant. By using the DARE system for visualizing the document space of generated document sets, it becomes clear that low levels of indexing exhaustivity and shallow term distributions, that a more dense document space results, whereas high exhaustivity and steep term distributions produce a more diffuse document space. With increases in both indexing factors, differences in the document space become smaller, making the influence of each factor less significant. The primary implication of the change in document space density for information retrieval influenced by the indexing characteristics is the retrieval system sensitivity, where the more diffuse space permits larger scale distinctions to be made among documents. Future research directions include the examination of the impact of indexing characteristics on document sets using the angle-based measure and a more detailed investigation of the larger changes in the document space observed at low exhaustivity and shallower distributions.

References

Benford, S., Greenhalgh, C., Snowdon, D., Ingram, R., & Knox, I. (1995). VR-VIBE: A virtual environment for co-operative information retrieval. Eurographics 95, 123–134.

Bennett, J.M. (1975). Storage design for information retrieval: Scarrott's conjecture and Zipf's Law. In E. Gelenb & D. Poitier (Eds.), International computing symposium (pp. 233–237). Amsterdam: North Holland.

Bird, P.R. (1974). The distribution of indexing depth in documentation systems. Journal of Documentation, 30(4), 381–390.

Burnett, J.E., Cooper, D., Lynch, M.F., Willett, P., & Wycherley, M. (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(3), 197–206.

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

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

Fekete, J.D., & Dufournaud, N. (2000). Compus: Visualization and analysis of structured documents for understanding social life in the 16th century. In Proceedings of the Fifth ACM Conference on Digital Libraries (pp. 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 (pp. 59–66). Denver, CO: ACM.

Helfman, J.I. (1994). Similarity patterns in language. In Proceedings 1994 IEEE Symposium on visual languages (pp. 173–175). St. Louis, MO: IEEE.

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

Kim, H., & Korfhage, R. (1994). BIRD: Browsing interface for the retrieval of documents. In Proceedings 1994 IEEE Symposium on visual languages (pp. 176–177). St. Louis, MO: IEEE.

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

Liu, Y.H., Dantzig, P., Sachs, M., Corey, J.T., Hinnebusch, M.T., Damashek, M., & Cohen, J. (2000). Visualizing document classification: A search aid for the digital library. Journal of the American Society for Information Science, 51(3), 216–227.

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.

Nuchprayoon, A., & Korfhage, R.R. (1994). GUIDO, A visual tool for retrieving documents. In Proceedings 1994 IEEE Computer Society Workshop on Visual Languages (pp. 64–71). St. Louis, MO: IEEE.

Olsen, K.A., & Korfhage, R.R. (1994). Desktop visualization. In Proceedings 1994 IEEE Symposium on Visual Languages (pp. 239–244). St. Louis, MO: IEEE.

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

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

Small, H. (1999). Visualizing science by citation mapping. Journal of the American Society for Information Science, 50(9), 799–813.

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.

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.

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(4), 296–305.

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

Wolfram, D. (1992b). Applying informetric characteristics of databases to IR system design, Part II: simulation comparisons. Information Processing & Management, 28(1), 135–151.

Young, D., & Shneiderman, B. (1993). A graphical filter/flow representation of Boolean queries: A prototype implementation and evaluation. Journal of the American Society for Information Science, 44(6), 327–339.

Zhang, J. (2000). A visual information retrieval tool. In Proceedings of the 63 Annual Meeting of the American Society for Information Science (pp. 248–257). Medford, NJ: Information Today, Inc.

Zhang, J. (2001). TOFIR: A tool of facilitating information retrieval-Introducing a visual retrieval model. Information Processing & Management, 37(4), 639–657.

Zhang, J., & Korfhage, R. (1999). DARE: Distance and angle retrieval environment: A tale of the two measures. Journal of the American Society for Information Science, 50(9), 779–787.

Zhang, J., & Wolfram, D. (2001). Visualization of term discrimination analysis. Journal of the American Society for Information Science and Technology, 52(8), 615–627.