The Influence of Indexing Practices and Weighting Algorithms on Document Spaces

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

Index modeling and computer simulation techniques are used to examine the influence of indexing frequency distributions, indexing exhaustivity distributions, and three weighting methods on hypothetical document spaces in a vector-based information retrieval (IR) system. The way documents are indexed plays an important role in retrieval. The authors demonstrate the influence of different indexing characteristics on document space density (DSD) changes and document space discriminative capacity for IR. Document environments that contain a relatively higher percentage of infrequently occurring terms provide lower density outcomes than do environments where a higher percentage of frequently occurring terms exists. Different indexing exhaustivity levels, however, have little influence on the document space densities. A weighting algorithm that favors higher weights for infrequently occurring terms results in the lowest overall document space densities, which allows documents to be more readily differentiated from one another. This in turn can positively influence IR. The authors also discuss the influence on outcomes using two methods of normalization of term weights (i.e., means and ranges) for the different weighting methods.

Introduction

Term weighting algorithms play an important role in information retrieval (IR). They are the foundation of automatic indexing, automatic abstracting, document clustering, and other automatic text processing tasks. Many term weighting algorithms have been developed over the years. To compare two term weighting algorithms or evaluate the effectiveness of an algorithm, researchers traditionally examine their effectiveness and efficiency on retrieval or use. More specifically, a term weighting algorithm to assign weights to keywords extracted from a document collection is selected, then the accuracy of these assignments is judged by people. Without a doubt, this kind of effectiveness and efficiency evaluation is important and indispensable. But as a complementary method, evaluating weighting algorithms from an internal database structure perspective, which investigates both inter-term-weight and intra-term-weight distributions, also is important.

Indexing approaches used in IR systems also play a vital role in the retrieval process (Salton, 1975). Subject searching for documents based on subject terms are guided by two indexing characteristics: the depth of indexing (i.e., exhaustivity), representing the number of terms used to index a document, and the preciseness of the indexing terms (i.e., specificity), representing the concentration and scatter of index terms within and across a document set. The less frequently a term appears in a document set, the more precise the term is said to be. The retrieval process and retrieval decisions made may be influenced by relying on different levels of exhaustivity and specificity along with various term weighting methods that provide more importance (or greater weight) to selected index terms based on their frequency of occurrence across documents or within documents.

Using an exploratory approach involving IR data modeling, visualization, and computer simulation, we investigated the impact of indexing changes and weighting methods on document space densities in a vector-based visual retrieval environment. Although the characteristics of document spaces have been studied for many years, to date, the influence of indexing characteristics and weighting methods on these spaces has not been widely explored. In an earlier study, we found that changes in indexing characteristics can affect the document space (Wolfram & Zhang, 2001). Previous research also has recognized that document space discriminative capacity changes are associated with retrieval effectiveness such that the characteristics of document spaces may affect retrieval performance. Specifically, the increase (or decrease) of a document space's discriminative capacity has a positive (or negative) effect on IR (Korfhage, 1997; Salton, 1975). A low DSD is associated with a good document space discrimination capacity. The converse applies to high document space densities, where the ability to distinguish documents from one another is reduced. Thus, in general, any indexing characteristic that can lead to an increase in the DSD will

Received April 3, 2006; revised June 6, 2006, December 20, 2006, February 16, 2007; accepted February 16, 2007

 ● Published online 1 October 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20688

result in a decrease of the document space discrimination capacity, and therefore decrease the retrieval effectiveness of the system.

In the present study, we investigated how changes in indexing policy as reflected in the index term frequencies (i.e., frequency distribution), the number of terms assigned to documents (i.e., indexing exhaustivity), and the term weighting algorithm used influence the document space discriminative capacity and density. If these relationships are clearly identified, one can predict what the impact will be on IR from a systems-centered perspective. Based on DSD changes caused by indexing changes, conclusions may be drawn regarding the impact of these characteristics and the term weighting schemes on overall retrieval. Note that the investigators recognize the importance of the user in determining retrieval effectiveness; however, the focus of the present investigation was to study system dynamics that exist independent of the user. User studies may be applied as follow-up to the present proposed research to validate findings.

The primary research question guiding this research is: What effect do different weighting approaches based on frequencies of occurrence have on document space densities in different indexing environments?

Previous Studies

This research intersects topics related to information visualization, indexing, and IR modeling. The core of this study is an exploration of several weighting methods on document spaces. Term weighting or automatic indexing is fundamental, essential, and vital for automatic information processing and IR. It is not surprising that considerable research has been published on different approaches to term weighting. The quality of text and ambiguity of natural language processing, however, have made progress in this area challenging (Wacholder, Evans, & Klavans, 2001).

The Inverse Document Frequency (*IDF*) algorithm of term weighting was introduced by Spark Jones in 1972. It has since been widely used in IR. The theoretical explanation and justification for the *IDF* algorithm continue to attract the attention of researchers in relation to the definition of the appropriate space for the probabilities concerned. Shannon and Weaver's information theory was used to interpret *IDF* as a particular kind of probabilistic function: a measure of the information amount possessed by a keyword (Robertson, 2004). Another probabilistic approach was presented for the same reason: Mutual information between term occurrence and relevance as a natural and useful measure of term quality is correlated with document frequency to derive a theoretical interpretation for the *IDF* algorithm (Greiff, 1998). Feature quantity, based on an information theoretic perspective of co-occurrence events between terms and documents, was defined to describe a quantitative representation of specificity. It maintains a good correspondence with term-frequency (*TF-IDF*)-like measures (Aizawa, 2000), which incorporate TFs within documents along with *IDF* data.

Bayesian inference networks have been applied to automatic indexing as well. They can offer a sound inference mechanism, integrate probabilistic concepts and theories, and use a graphic representation of the incorporated interdependencies (Tzeras & Hartmann, 1993). Term overlap and embedding were exploited to generate a substantial qualitative and quantitative improvement in automatic indexing through concept combination (Jacquemin, 1998). Jacquemin, Klavans, and Tzoukermann (1997) used a seed term list coupled with derivational morphology to achieve greater coverage of multiword terms for indexing. An automatic indexing algorithm based on a Hopfield network was presented. Concepts extracted from a document were used as the input pattern to a concept space represented as a Hopfield network. From this, a network parallel spreading activation process yielded a set of concepts (Chung, Pottenger, & Schatz, 1998).

Wacholder et al. (2001) introduced an evaluation standard that included three criteria to assess the quality of indexing terms: coherence, thoroughness of the coverage of document content, and usefulness of the index terms. Index terms, derived without resource lexicons or other kinds of domain-specific information, were evaluated using three shallow processing methods; however, most of the evaluations for weighting algorithms were not from a system database structure perspective.

The influence of indexing characteristics on retrieval efficiency and effectiveness has been studied from different perspectives over the past several decades. 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 that 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 datasets using hashing, truncation, and n-gram encoding techniques. Wolfram (1992a, 1992b) examined how different indexing exhaustivity and specificity patterns influenced retrieval speed and storage requirements for three different types of IR database environments and six different file structures. The study revealed that some file structures were better suited to different types of indexing environments based on a series of simulation studies.

We have previously studied the influence of indexing exhaustivity and specificity on indexing environments (Wolfram & Zhang 2001, 2002) using a *TF-IDF* term weighting scheme, but did not compare different weighting methods. Results of document space mappings of simulated indexing environments demonstrated that higher specificity environments resulted in document spaces with lower densities, which improve document differentiation during retrieval. The Wolfram and Zhang (2001) study examined the impact of singly occurring and frequently occurring

DOI: 10.1002/asi

terms on document space characteristics. The removal of tokens (i.e., specific occurrences) of singly occurring terms from documents notably influenced the density of the resulting document space by producing higher density spaces, thereby decreasing the ability to differentiate documents from one another. The removal of an equivalent number of tokens from frequently occurring terms had little influence on the DSD. This becomes clearer in a visualized environment.

By combining index modeling and information simulation methods, this study examines term weighting algorithms from a database internal structure perspective and presents a novel methodology for term weighting algorithm evaluation.

Method

Descriptor frequency and indexing exhaustivity data for 1,000 bibliographic records dealing with the topic of information technology were initially extracted from the ERIC bibliographic database. Descriptors were used instead of keywords because the former provides a more accurate subject description of document content. Although the dataset is small compared to today's gigabyte data sources, the investigators wished to keep the dataset size manageable for processing purposes using an IR environment, the *Distance Angle Retrieval Environment* (*DARE*; discussed later), developed by Zhang and Korfhage (1999).

Frequency distributions for both exhaustivity and descriptor frequency data were first tabulated, resulting in observed distributions. The observed distributions were supplemented with Zipfian and unimodal hypothetical distributions that changed the characteristics of the observed datasets and the mean values of the frequency distributions. In essence, these modified distributions represent different indexing models with lower and higher levels of indexing exhaustivity and specificity. Figure 1 shows the exhaustivity distributions used in the study representing low ($M = 7$ terms per document), observed ($M = 11$ terms per document), and high ($M = 15$ terms per document) levels of exhaustivity. The hypothetical low and high exhaustivity distributions were based on negative binomial models, which have been shown to be representative of exhaustivity in actual environments (Bird, 1974; Nelson & Tague, 1985).

Figure 2 shows the distribution of descriptor occurrences across the document set; that is, the number of documents that contain a given descriptor. Hypothetical distributions were based on Zipfian models, which previous research has shown to be characteristic of retrieval system indexes (Nelson, 1989; Wolfram, 1992a). Two hypothetical distributions were developed: the first used a higher level of specificity (i.e., a lower mean frequency per descriptor with a more steeply descending distribution), and the other used a lower level of specificity (i.e., a higher mean frequency with a more shallowly descending distribution, with a greater number of higher frequency descriptors). These hypothetical distributions were achieved by altering the Zipfian model parameters such that mean frequencies for each distribution represented three, six, and nine occurrences per generated descriptor, respectively.

The three levels of indexing exhaustivity (i.e., low, observed, high) and three levels of indexing specificity (i.e., shallow, observed, steep) provide nine combinations of indexing characteristics, representing very thorough, precise

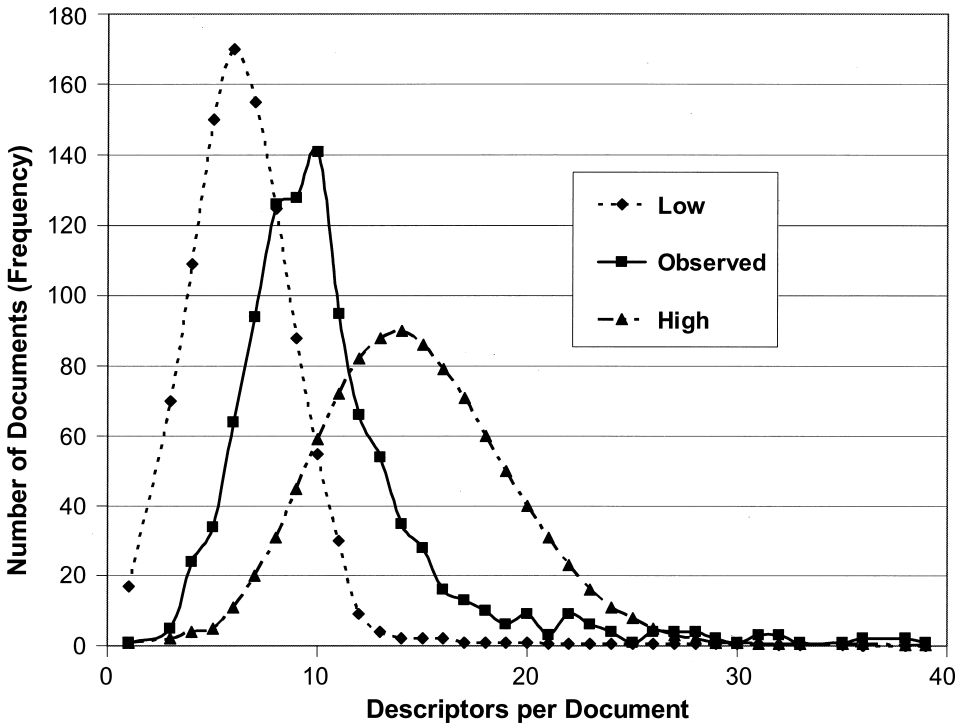

FIG. 1. Descriptor exhaustivity distributions.

DOI: 10.1002/asi

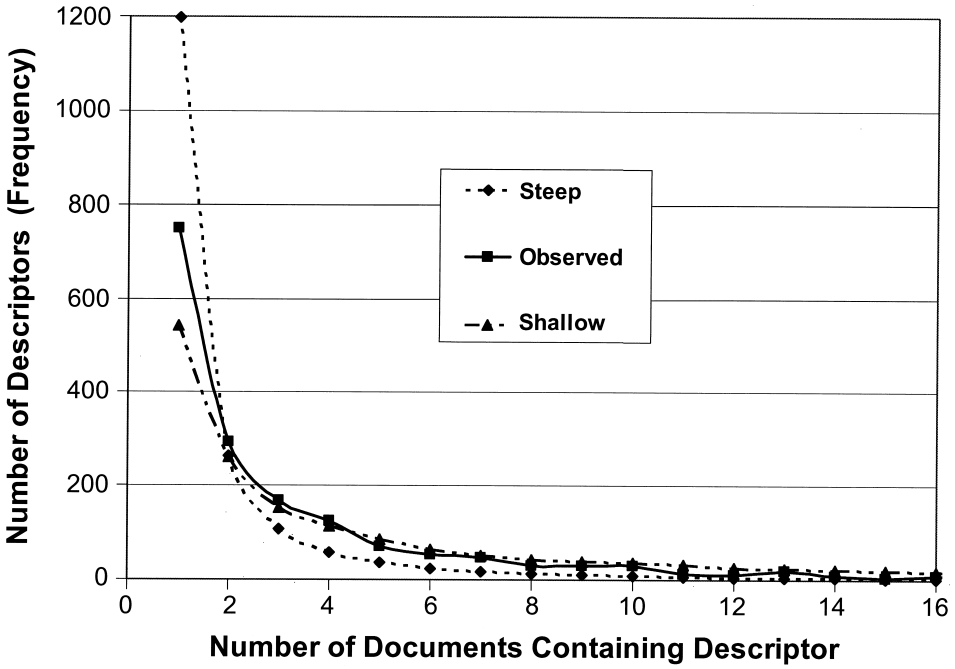

FIG. 2. Descriptor frequency distributions.

indexing (i.e., high exhaustivity and specificity), and parsimonious indexing (i.e., low exhaustivity and specificity) characteristics. Different indexing characteristics will produce different numbers of index-term types across document sets. As in Wolfram and Zhang (2002), for each simulation run, exhaustivity values were generated first for each document, with descriptor frequencies generated for each document term. A descriptor token was generated based on the selected Zipfian distribution. Exhaustivity levels and descriptor tokens for the observed distributions also were randomly generated to allow a comparison with the hypothetical exhaustivity and descriptor frequency distributions. Tallies were kept of how frequently distinct descriptors were selected to ensure that distribution characteristics were maintained. The program continued to generate additional documents until the number of distinct descriptors generated approached the total observed number of descriptors. By keeping the number of distinct descriptors relatively constant, with only minor variations that result from a small number of additional descriptors generated for the final document, the spaces provide a relatively uniform dimensional environment in which to compare outcomes. By fixing the number of descriptors across the different exhaustivity and indexing levels, at least one characteristic of the generated document sets must be allowed to vary. For this study, the number of documents varied depending on the exhaustivity level and descriptor frequency distribution. For example, to generate an equal number of descriptor types for low- and high-exhaustivity environments, the low-exhaustivity environment must generate a larger number of documents. Similarly, with shallow and steep descriptor frequency distributions, a larger number of documents must be generated for the shallow distribution.

We acknowledge the existence of term (or descriptor) dependence on term co-occurrence, where the presence of one term results in a higher incidence of co-occurrence with another term than dictated by chance alone; however, the models used assume independence of descriptor co-occurrence. In varying the exhaustivity and indexing specificity levels, the influence of this dependence in each environment is unknown. In addition, the vocabulary generated was hypothetical, and the strength of co-occurrence relationship has been shown to be significant for only a small percentage of term pairs (Jacquesson & Schieber, 1973).

The indexing characteristics define how descriptors are assigned to documents, but not their significance. Weights assigned to terms or descriptors may be based on their frequencies across the document set. Classic *IDF* weight assignment provides higher weights to lower frequency terms, thereby resulting in an inverse relationship between frequency and weight (Salton, 1975). One also could assign weights based on a direct relationship, where higher frequencies across the document set result in a higher assigned weight (assuming traditional stopwords have been filtered). Descriptor weights for each occurrence also were assigned based on observed and hypothetical weight distributions. We wanted to introduce additional variability of weights based on intradocument importance of each descriptor. These values were not available through ERIC but were modeled on other observed data. Through analysis of index term weight characteristics used in a previous study, Zhang and Nguyen (2005) found that different weighting algorithms produced different probability distributions of average weights for indexable terms of varying frequencies.

An additional hypothetical weighting scheme that assumes uniform assignment of weights on average is included

DOI: 10.1002/asi

in the study for comparison. With a uniform scheme, there is no relationship between term frequency and the average assigned weight. The general distribution of mean term weights for each of these scenarios appears in Figure 3. Because weight assignment values are variable in actual systems due to the relative significance of a given descriptor within a document when compared to another document, a method for randomly assigning a weight around the mean weights for each occurrence is needed. In this case, generated descriptor weights for each descriptor frequency across the document set were based on a lognormal distribution of a given mean (determined by the descriptor frequency) and standard deviation (fixed at 0.5 across all descriptor frequencies), resulting in a range of weights for each descriptor frequency across the document set. The use of a lognormal distribution, instead of a normal distribution of weights, more closely models the skewed distribution of term weights for each frequency as observed in the dataset used by Zhang and Nguyen (2005).

The generated documents based on the descriptor frequency and exhaustivity distributions were projected onto the *DARE* (Zhang, 2000; Zhang & Korfhage, 1999) visual IR environment. *DARE* 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. *DARE* was originally developed for visualized IR. Later, it was used as a testing platform for term discrimination analysis (Zhang & Wolfram, 2001) because it can display spatial-distance characteristics of a document.

Briefly, if two reference points in a vector-based document space are defined, these two reference points can determine a reference line in the document vector space. Using this reference line, one can define any documents in the vector space by two parameters. One is the visual distance, defined as the distance from that document to one of the two reference points. The other is the visual angle, defined as the angle formed by the document and the reference point against the reference line. As long as the two reference points and the reference line are clearly defined, these two parameters are always available for a document (i.e., the two parameters of a document that will be employed for the *DARE* visual space construction). Based on the two parameters of a document, the document in a high dimensional space can be easily projected onto a low two-dimensional visual space. The visual angle and visual distance are assigned to the *x* axis and *y* axis, respectively, of the visual space. The legitimate value of the *x* axis of the visual area is from zero to π because a document always is symmetrical against the reference line. Because there is no limitation for the visual distance of a document, the *y* axis can range theoretically from zero to infinity. Although not used for its visualization ability for this study, *DARE* was used to calculate the DSD of the generated indexing environments.

The resulting document space was evaluated by comparing the average document distances to the document set centroid (or average document vector), providing an indication of the DSD. The density serves as a normalized method for comparing the document space characteristics of each environment studied.

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

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

$Dist(D_i,C)$ represents the Euclidean distance between D_i and Centroid C, n represents the total number of documents in the document set, and $\lceil x \rceil$ represents the value of x rounded up. Note that distance values were never below 1 for the generated data. By dividing by the total number of documents in the generated document set, this takes into account the differences in the generated number of documents across the different combinations of indexing characteristics.

To permit comparisons of the resulting document spaces across the term weighting methods, which by their formulaic representation produce different absolute values for term weights, the distribution of these weights also must be normalized. Two approaches are used for this purpose:

- *Equalization of the mean descriptor weights.* Descriptor weights within a generated document set were divided by the mean of all generated values. This method allowed the

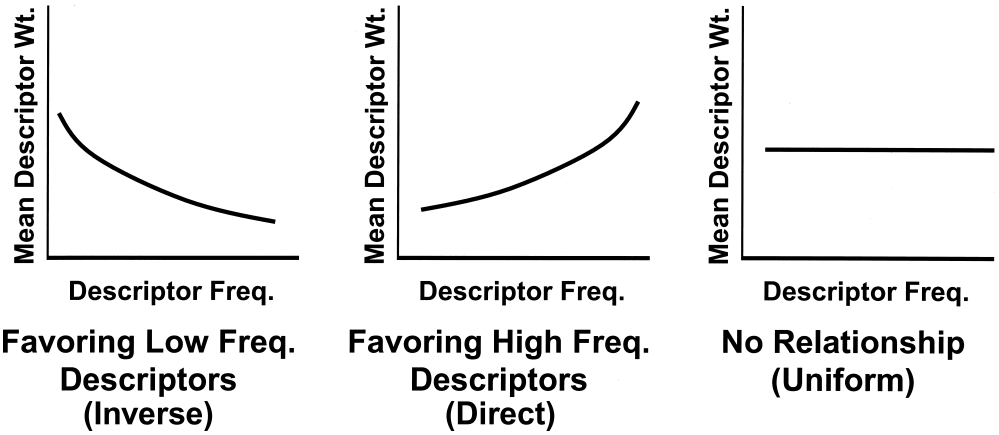

FIG. 3. Distribution of mean descriptor weights by weighting method.

DOI: 10.1002/asi

ranges of values to vary so that each descriptor weighting method distribution resulted in a mean weight of 1.0 across all generated values, but with a range of normalized values between zero and an open-ended maximum.

- *Equalization of the range of descriptor weights.* The maximum and minimum values across all occurrences (but not for each descriptor frequency) were equalized for each weighting method; however, with equalized ranges, the mean descriptor weight values for each method will vary. For example, in one simulation run, the raw outputs for each weighting method resulted in values between 0.05 and 9.00 for the inverse relationship, 0.13 and 13.37 for the direct relationship, and 0.14 and 10.05 for the uniform relationship. These were transformed by projecting the raw values onto a standardized range of 0.1 to 50 to equalize the spread of values for each run and simulation combination.

To reduce the influence of the variability of the outcomes of individual simulation runs, five runs for each combination of characteristics were conducted, with the mean outcomes for each combination reported.

Note that these equalization methods should not be confused with traditional normalization used in IR document generation or retrieval. The methods in the present study are performed after the generation of descriptors, and are used solely to permit a level comparison between each method and do not affect the relationships among descriptors within each generated set.

Results

Plots of the average DSD values for each combination and term weighting scheme for equalized ranges and mean DSDs appear in Figures 4 and 5, respectively. Mean values for exhaustivity levels and descriptor frequencies for each generated document set were tabulated to ensure the simulated sets closely followed the modeled indexing characteristics. The mean generated values across the simulation runs for each combination never varied by more than 10% of the targeted mean for either exhaustivity or descriptor frequency distributions. The observed variability was due to the randomness of the generated values within each simulated dataset, which influenced the mean of the outcomes.

The normalization method used clearly affects the distribution of documents within the document space. Higher weights assigned to lower frequency descriptors result in the least dense spaces in both cases. The opposite is observed for weighting favoring high-frequency descriptors when using mean DSD normalization. As observed in Wolfram and Zhang (2002), steeper frequency distributions (i.e., higher specificity) result in lower density spaces, with indexing exhaustivity playing a lesser role. The differences in DSD across the different weighting methods also become smaller as indexing specificity increases. Variations in outcomes for the different indexing exhaustivity levels are undoubtedly due to the generated values for the simulation runs, and are not as strong as the overall pattern that emerges as a result of the indexing specificity.

DSD values for the inverse weighting relationship are the lowest for each indexing characteristic and each normalization method, but are less affected by indexing specificity or exhaustivity for the normalized means, based on the lower range of DSD values observed across the different indexing characteristics. The direct weighting relationship results in the highest density spaces for normalized means and second highest for normalized ranges. This difference is likely due to the range of values encountered in each method. Documents are more easily distinguished from one another by having larger differences in weights for each descriptor frequency. Both the inverse and direct weight methods produce larger ranges overall based on their distribution of possible

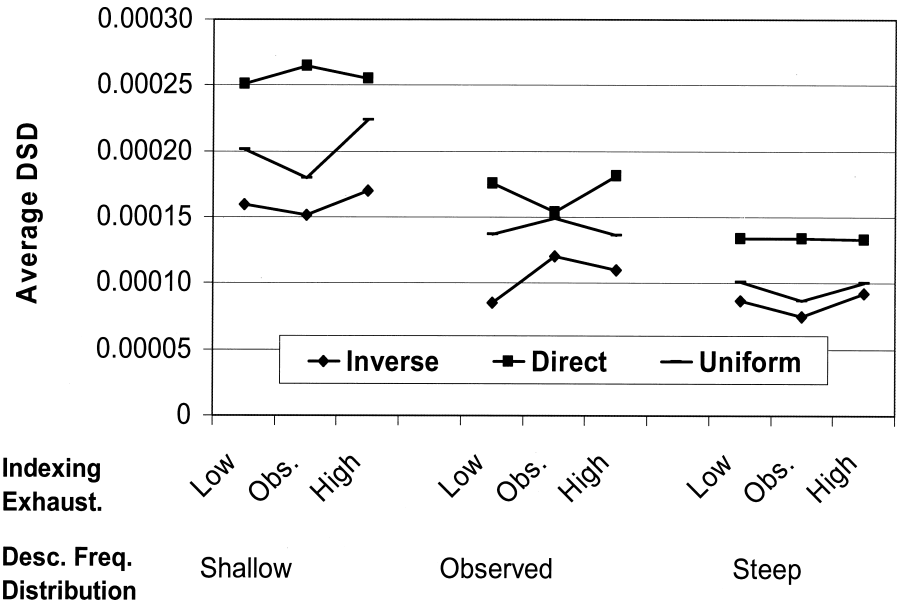

FIG. 4. DSD outcomes using normalized ranges for term weights.

DOI: 10.1002/asi

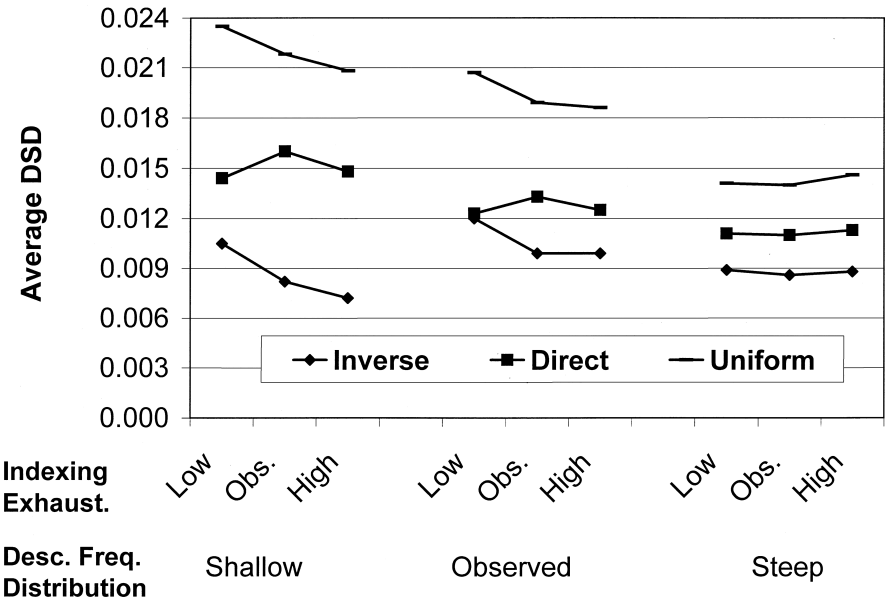

FIG. 5. DSD outcomes using normalized means for term weights.

values for different term frequencies. By equalizing the ranges for each of the three methods, the uniform approach benefits because the range of values generated for a given descriptor frequency is larger than that for the equalized mean method. The outcome is a more diffusely arranged document space. The direct method does not benefit as much from equalized ranges because there are fewer occurrences produced by high-frequency descriptors that result in higher weights. With fewer occurrences of more extreme weights, densities within the document space increase. The inverse method, conversely, produces higher weights for the greater number of occurrences produced by the low-frequency descriptors, providing more contrast between documents and lower density spaces.

Discussion

Comparison of different weighting methods cannot be performed objectively unless outcomes are normalized in some way. Without a method to normalize the term weight values, outcomes for each method cannot be compared directly. The type of normalization used, however, affects the performance of different measures so that one algorithm may perform better in one environment and worse in another, as was observed for the direct relationship in weight assignment. The normalization methods used were applied after descriptors had been assigned to documents and, therefore, were independent of traditional weight normalization used in IR. The inverse relationship resulted in the least dense spaces regardless of the normalization method. The differences in performance of each weighting method as a result of the normalization method are notable for the uniform approach.

Should one rely on equalized means or ranges to permit comparison of the outcomes across different weighting methods? Normalized means produce roughly equivalent variability in weights for descriptors of given sizes, but not for the overall range of values across all descriptor sizes, which may be small or large while still having the same mean. Normalized ranges can produce a larger or smaller amount of variability in the weights depending on the range selected. In the present study, the range of values was standardized to be larger than the range observed for standardized means. An advantage of range equalization is that the transformation does not alter relationships among data points. Because document space characteristics are determined by distances based on descriptor weights, normalizing the ranges ensures that values are of the same scale while maintaining ordinal relationships between the calculated weights. The equalized range used resulted in a comparatively much less dense document space than that using normalized means.

This study demonstrates how indexing specificity more strongly influences document space characteristics than does indexing exhaustivity, regardless of the weighting assignment method used. The lack of a notable regularity in DSD change stemming from exhaustivity indicates that it does not play as strong a role as does specificity in determining the density of the document space. The declining differences in DSD between the three methods with increased specificity demonstrate that the significance of the method used decreases with higher specificity. With more unique descriptors used to describe document content, the way one identifies the relative importance of a descriptor becomes less important. This finding underscores the importance of including descriptors of low frequency. Because they each serve as access points for only a small number of documents, they may be viewed from a system design perspective as unimportant for retrieval purposes while occupying valuable space in database indexes from a storage perspective. As noted in Korfhage (1997) and Salton (1975), low DSD is associated with a good document space discrimination capacity.

DOI: 10.1002/asi

A weighting scheme based on inverse document frequency resulted in the lowest density spaces, even when intradocument differences in assigned descriptor weights were taken into account. This suggests a better environment that is conducive to more effective retrieval.

Limitations of the study arise from constraints placed on data to control for confounding factors that may influence outcomes. First, by relying on a comparatively small descriptor and document set, generalizability to larger database environments is limited. Computational burden resulting from the high dimensionality of the datasets becomes an issue with large, commercial systems, and is not compatible with the *DARE* environment. In generating descriptor/document relationships, the assumption of indexing independence was made. The close relationship that small numbers of terms have with one another that cause them to co-occur with one another in documents has long been known (Jacquesson & Schieber, 1973). The generation of these relationships and term affinities in a hypothetical environment, particularly when indexing characteristics such as term frequency distributions, however, becomes speculative. Normalization of results is necessary to make outcomes comparable. The normalization method clearly affects outcomes and the conclusions drawn. By relying on two such methods, outcomes can be compared. The distribution of average weights for each weighting scheme may vary in different environments, but the general characteristics of favoring terms of given sizes with higher average weights remain the same.

Finally, a lower DSD by itself does not definitively determine the retrieval performance; however, it does influence the environment in which retrieval takes place. The more widely dispersed the documents are, the more easily they are differentiated. One may argue that this is resolved through scaling of any document space; however, if two documents are superimposed over one another in a document space due to limited indexing practice, no amount of scaling will help to differentiate them.

Conclusions

The current study contributes to the understanding of weighting algorithms and provides a unique way to evaluate weighting algorithms from a system perspective, and more specifically, this research contributes in three areas. First, it reveals how given frequency distributions and levels of indexing exhaustivity impact the DSD. A higher density space makes it more difficult to distinguish one document from another, which may reduce the system effectiveness. Second, the study demonstrates how different weighting approaches can influence the characteristics of the document space. Finally, the method used allows investigators to essentially experiment with different indexing environments to determine possible influences on document characteristics, which may influence retrieval performance.

By comparing the outcomes using three hypothetical weighting patterns, the authors conclude that the lowest density spaces, which ultimately are conducive to more effective retrieval, result when the distribution of assigned term weights is inversely proportional to the term/descriptor frequency [similar to Salton's (1975) classic term weighting algorithm]. The findings have implications for automatic indexing practice within bibliographic database environments by providing insights into how different term/descriptor frequency and indexing exhaustivity distributions can influence document spaces. Different weighting methods may influence retrieval effectiveness by retrieving (or not retrieving) the most relevant documents to a given query. Ideally, a high-specificity environment is desirable, but the rate of addition of new terms/descriptors as the database grows declines as one approaches saturation of a finite vocabulary, thereby making high specificity in large database environments difficult to achieve.

This research provides a foundation for a future study that will examine the impact of indexing characteristics on the complementary concept of a *term space*. By identifying index terms whose presence or absence most significantly influences the term space and document space, one may identify which terms are the most effective for retrieval purposes from a systems-centered perspective. The influence of term weight value ranges also merits further investigation. In addition, future research will compare retrieval performance based on different term weighting methods by conducting a user study that incorporates these indexing methods. Knowledge of this outcome also may be used to derive a term weighting algorithm that capitalizes on the index frequency distribution characteristics of a given database environment along with the term weight range for more effective retrieval.

Acknowledgment

We thank the anonymous reviewers for their constructive comments.

References

Aizawa, A. (2000). The feature quantity: An information theoretic perspective of TF-IDF-like measures. Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 23, 104–111. New York: ACM Press.

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.

Chung, Y.M., Pottenger, W.M., & Schatz, B.R. (1998). Automatic subject indexing using an associative neural network. In Proceedings of the 3rd ACM Conference on Digital Libraries (pp. 59–68). New York: ACM Press.

Greiff, W.R. (1998). A theory of term weighting based exploratory data analysis. In Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 21, 11–19. New York: ACM Press.

Jacquemin, C. (1998). Improving automatic indexing through concept combination and term enrichment. Proceedings of the 36th annual meeting of the Association for Computational Linguistics (pp. 595–599). Montreal, Canada.

DOI: 10.1002/asi

Jacquemin, C., Klavans, J.L., & Tzoukermann, E. (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. Proceedings of the 8th Conference of the European Chapter of the Association for Computational Linguistics (pp. 24–31). Madrid, Spain.

Jacquesson, A., & Schieber, W.D. (1973). Term association analysis on a large file of bibliographic data using a highly-controlled indexing vocabulary. Information Storage and Retrieval, 9, 85–94.

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

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.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation, 60, 503–520.

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

Spark Jones, K. (1972). A statistical interpretation of term importance in automatic indexing. Journal of Documentation, 28, 11–21.

Tzeras, K., & Hartmann, S. (1993). Automatic indexing based on Bayesian inference networks. Proceedings of the 16th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 16, 22–35. New York: ACM Press.

Wacholder, N., Evans, D.K., & Klavans, J.L. (2001). Automatic identification and organization of index terms for interactive browsing. Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 126–134). New York: ACM Press.

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.

Wolfram, D., & Zhang, J. (2001). The impact of term indexing characteristics on a document space. Canadian Journal of Information and Library Science, 26(4), 21–35.

Wolfram, D., & Zhang, J. (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(11), 943–952.

Zhang, J. (2000). A visual information retrieval tool. Proceedings of the 63rd annual meeting of the American Society for Information Science (pp. 248–257). Chicago, IL.

Zhang, J. (2001). TOFIR: A tool of facilitating information retrieval—Introduce 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., & Nguyen, T.N. (2005). A new term significance weighting approach. Journal of Intelligent Information Systems, 24(1), 61–85.

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

DOI: 10.1002/asi