Visualization of Term Discrimination Analysis

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

A visual term discrimination value analysis method is introduced using a document density space within a distance–angle-based visual information retrieval environment. The term discrimination capacity is analyzed using the comparison of the distance and angle-based visual representations with and without a specified term, thereby allowing the user to see the impact of the term on individual documents within the density space. Next, the concept of a "term density space" is introduced for term discrimination analysis. Using this concept, a term discrimination capacity for distinguishing one term from others in the space can also be visualized within the visual space. Applications of these methods facilitate more effective assignment of term weights to index terms within documents and may assist searchers in the selection of search terms.

Introduction

Information visualization has become an important area of investigation within a number of disciplines in the physical and social sciences. Visualization borrows from many areas of study such as psychology, human–computer interaction, data mining, imaging, and graphic design. Research in information visualization focuses on helping people to visualize abstract or conceptual information by reducing its complexity. Fundamental problems for visualization research include the discovery of special presentation approaches for displaying hidden information and the understanding of how visual presentations are interpretable and meaningful for special analytical tasks confronted in everyday life. Due to the intangible characteristics of abstract information, people cannot observe, perceive, and analyze information as a tangible object with obvious physical representation.

Visualization techniques have also been applied to the information retrieval environment where documents and their relationships may be considered an abstract information space. This space may potentially contain thousands of documents. It is clear that there exist relationships among these documents or even indexing terms in the document collection, but they are invisible because of the high database dimensionality.

Visualization of information transforms the unseen internal semantic representations of information retrieval systems into visible geometric displays, and demonstrates internal processes for users. Simply put, visualization offers a method for seeing the unseen. Visualization provides a number of benefits (Zhang, 1999): (1) The understanding of internal relationships among documents helps users to make decisions in judging relevant documents in a search; (2) a transparent process makes the search and analysis easier and more effective; (3) a visual environment presents richer information for users; (4) visualization has the potential to provide new methods of information processing; (5) visualization brings the recognition capacities of human beings to bear in either the discovery or display of information. In short, information visualization in this domain can enable users to understand information better, to get information more quickly, and to make more reasonable decisions about document relevance.

How and what are visualized in a visualization space are fundamental and crucial questions that system designers must confront. Answers to these questions will determine the features and functions of an information visualization tool. Some attributes of an object, such as a document, can be extracted to build a visual space, which is a key part of an information visualization tool. Visualized objects in a visual space can be documents such as full texts (Fekete & Dufournaud, 2000; Hearst, 1995; Helfman, 1994), document flows (Young & Shneiderman, 1993), document sets (Borner, 2000; Brooks & Campbell, 1999; Furnas & Rauch, 1998; Michard, 1982; Shneiderman, Feldman, Rose, & Grau, 2000), simple document icons (Heath et al., 1995; Kim & Korfhage, 1994; Nuchprayoon & Korfhage, 1994; Olsen & Korfhage, 1994; Wise, 1999; Zhang & Korfhage, 1999); document citations (Small, 1999); and subjects or subject terms presenting documents (Fowler, Fowler, & Wilson, 1991; Lin, 1993).

Received August 21, 2000; Revised November 20, 2000; accepted December 12, 2000

Term discrimination analysis has been an important area in information retrieval research for several decades. Its primary application has been for automatic indexing of documents. Term discrimination addresses the capacity that allows a document to distinguish itself from others in a document collection. van Rijsbergen (1979) shows that there are two distinct ways to characterize a document: (a) representation without discrimination, and (b) discrimination without representation. Representation without discrimination characterizes a document just from its contents, ignoring the impact of other documents in the document collection. Discrimination without representation characterizes a document only from the perspective that it discriminates itself from other documents in the collection, regardless of its contents. Because each document in a document collection is not independent of other documents, and they share the same topics or related topics, the mutual impacts among the documents should be taken into consideration for this characterization. From this it is clear that neither should be ignored. Basically, computation of the discrimination capacity of a term involves determining an average similarity for all documents in the document collection by considering each pair of documents in that collection. Such a computation is an $O(N^2)$ process and, therefore, quite time consuming for a large document collection (Korfhage, 1997).

In the present study, the authors deal with an application of the traditional vector space model used in information retrieval. The concept of a "space density" for a document collection, introduced by Salton (1989), is used to describe the relationships among documents. Term discrimination values can be computed as the difference of the document space densities before and after a term assignment to a document collection. The space density can be obtained by employing the average similarity between all term pairs in the collection. Approaches to the similarity measures can be distance based or angle based. Other methods may also be used. Computational complexity of the approach is a major concern. To reduce computational complexity, the centroid of all terms in the document collection, representing the geographic center of all terms in the document space, replaces the average pairwise similarity between all term pairs. The computation for this strategy is reduced to an O(N) process. Other related research on improving these algorithms has been undertaken by Willett (1985), El-Hamdouchi and Willett (1988), and Biru, El-Hamdouchi, and Rees (1989).

After the discrimination value of a term is determined, it is usually used to modify the term weight to make it more reasonable, sound, and accurate, which is the primary motivation for term discrimination analysis. It has also been used to improve access efficiency to signature files (Chang, Lee, & Lee, 1989).

It is apparent that the weaknesses of the traditional approach to term discrimination value determinations are that: the procedures are complicated, making the corresponding term discrimination value computation timing consuming; the difference of the document space densities before and after a term assignment cannot be perceived by users; and, the overall change of the space densities is expressed only by a single term discrimination value.

Note that the document density space is an abstract concept that exists in the document collection but is invisible. Application of visualization techniques to term discrimination analysis may shed light on this issue. A visual display of the space densities rather than a single difference value would help users to understand the nature of the density space and their differences, presenting richer information for decision making in a visual environment. Although visualized objects in some visual tools discussed earlier are subjects or subject terms, there is no visual tool for term discrimination analysis, let alone one using the distance–angle-based approach to analyze the term discriminative capacity visually.

The purpose of this article is to demonstrate the application of the distance–angle-based visual tool to visualize the document (and term) density space changes in a document collection to facilitate selection of discriminating terms. Both document density space and indexing term density space are visualized for term discrimination capacity analysis. Traditionally, this has been done by analyzing term discrimination capacity via the document density spaces before and after a term assignment to a document collection. In addition to this traditional approach, we investigate the impact before and after an indexing term assignment to a document collection on the indexing term density space. Within the term density visual environment, relationships among terms are demonstrated, the two term distributions before and after a term assignment to the collection are compared, and ultimately the term discriminative capacity is analyzed based on the two density spaces.

The unique contributions of this study are that (a) the visualization technique is employed to analyze the document density space change, (b) a new concept called "term density space" is introduced and the corresponding space is visualized, and (c) the relationships between the document density space and the indexing term density space before and after a term assignment to a document collection are made. Visualization for term discriminative capacity analysis will not only enrich information visualization research, but also presents a unique way to perform term discrimination research. The present study represents the first part of a multistage investigation of applications of visualization for the design of more effective vector-based IR systems. These applications will allow more informed decisions to be made in indexing IR system contents, and ultimately provide searchers additional decision support in the selection of search terms during query formulation.

The Distance–Angle-Based Visual Tool

The distance–angle-based visual tool *DARE (Distance Angle Retrieval Environment)* (Zhang & Korfhage, 1999), is a two-dimensional visual retrieval tool, consisting of a

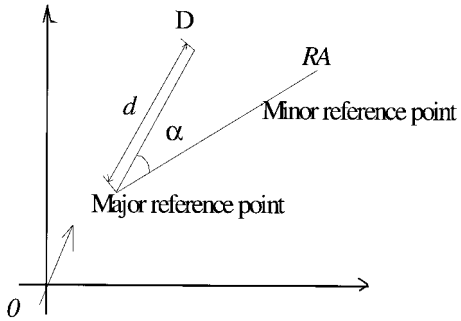

FIG. 1. A document in a document space.

graphical representation of the visual distance and the visual angle as the *X*-axis and *Y*-axis, respectively. Suppose there are two special points in a document vector space represented as point *I* and point *II*. These points constitute the transformation of a query submitted by the searcher. Using these points, the visual distance and the visual angle of a document (*D*) in a document vector space can be defined as the distance (*d*) from point *I* to the document (*D*), and the angle (α) formed by the document (*D*) and the point *II* against the point *I* respectively (see Fig. 1). Here, the points *I* and *II* are called the major reference point and minor reference point, respectively.

These two parameters are indispensable and fundamental for building the visual space. Based on the two parameters of a document, it can be accurately mapped onto a two-dimensional visual space whose *X*-axis and *Y*-axis are defined as the visual angle and visual distance, respectively. In other words, any document can be positioned in the visual space if the two reference points are selected and its two parameters are measured. The *X*-axis range of the visual area is defined from zero to π due to the symmetry of the visual angle against the reference axis (*RA*) generated by the two reference points and the simplicity of the visual space display. Because there is no limitation on a visual distance of a document, the *Y*-axis ranges from zero to infinity. The analysis suggests that the document projection area is an open rectangle area in the first quadrant of the visual space. Figures 1 and 2 show the case of a document *D* in a vector-based document space and its projected position in the distance–angle-based visual space, respectively.

The major and minor reference points can also be projected onto the visual space. Their positions are fixed relative to a document. The major reference point is always situated on the origin of the visual space due to the fact that its visual distance and visual angle are both always equal to zero. The minor reference point is always mapped onto the *Y*-axis due to the fact that its visual angle is equal to zero and the visual distance is the distance from the major reference point to the minor reference point.

Now let us address basic characteristics of the visual space and the document projection. A document distribution in the visual space is relative. In the *DARE* visual space the whole document distribution in the visual space is determined via the two reference points. It implies that changing the major and/or minor reference point(s) will result in a change to the document distribution in the visual space even if the real relationships among all documents in the document vector space stays the same. A document may be assigned to different visual angles and visual distances if the reference points for the projection change. It is obvious that the major reference point determines both the visual distance and the visual angle of a document while the minor reference point only affects its visual angle.

There are two basic document projection modes: one is the origin-based mode in which the origin of the document vector space is always assigned as the major reference point, while the other is the nonorigin-based mode in which the major reference point can be any other point in a document space but the origin.

Within the visual space, document icons near the *X*-axis are considered more relevant to the reference points in terms of the distance-based similarity measure, while those near the *Y*-axis are regarded as more relevant to the reference points in terms of the angle-based similarity measure. Document icons located around the origin of the visual space are relevant in terms of both similarity measures.

DARE was initially developed for visualization of document spaces for information retrieval. Several information retrieval models may be visualized and interpreted within *DARE* (e.g., Conjunction model, Disjunction model, Ellipse model). A sample interface of *DARE* appears in Figure 3. *DARE* can visualize not only the document density space, as discussed above, but also the indexing term density space. The difference between these two visualizations resides primarily in their database structures: the former employs a document-term vector as its primary data structure, while the latter utilizes a term-document vector. Both spaces are high dimensional, and objects in both are not easily visualized. The two vectors can be transformed from one to the other via a vector-reversing operation. The projection algorithm for both the document density visual space and the indexing term density visual space is the same. However, the visualized objects in the two visual density spaces are totally different. The visualized objects in the document visual space are documents, while the visualized objects in the indexing term visual space are indexing terms.

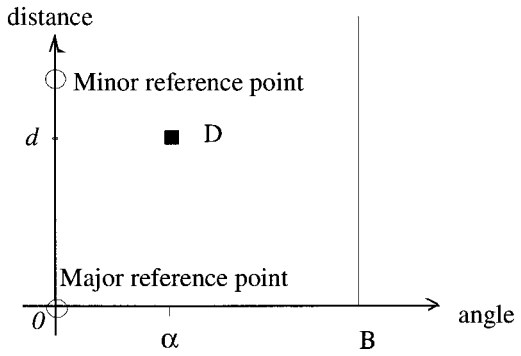

FIG. 2. A document in the visual space.

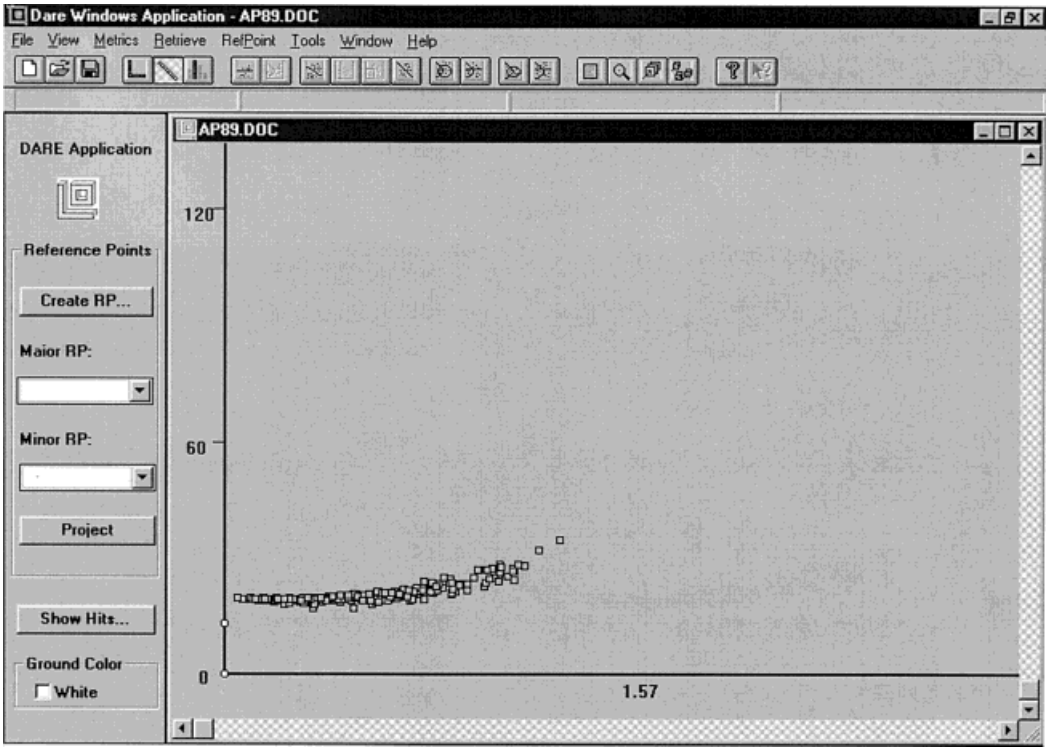

FIG. 3. The DARE interface.

Visualization of Term Discrimination

Document Density Space Analysis

Traditional term discrimination analysis is based on a document-term vector, where the document density space is computed and generated. The document-term vector 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.

The discrimination value of a term refers to the degree to which the term can distinguish documents in a document collection. The degree is obviously affected by not only the term itself but also other terms in the collection.

Average similarities of all documents both with and without a specified indexing term are compared to get its discrimination value. A more efficient document space density computation is based on the use of a dummy document, called the document centroid. 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:

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

where

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

Here, n is the number of all terms in the collection, and k is the number of nonzero elements in

$$\sum_{j=1}^{m} t_{ji}$$

The document centroid without term t_i can be expressed in Equation (4).

$$D_{ci} = (c_1, c_2, \ldots, c_{i-1}, 0, c_{i+1}, \ldots, c_n) \quad (4)$$

Because term t_i is extracted from the document vector, its corresponding value in the centroid D_{ci} is set to zero.

The discrimination value of term t_i is defined as:

$$\nu_i = a^x \left(\sum_{k=1}^{n} \text{Similarity}\,(D_k, D_{ci}) - \sum_{k=1}^{n} \text{Similarity}\,(D_k, D_c) \right) \tag{5}$$

In Equation (5), Similarity (*X, Y*) stands for the similarity between document *X* and document *Y*. The similarity measure can be either the distance-based or angle-based measure. The parameter *a* is a suitable constant. The discrimination value is equal to the difference between the average similarities without term t_i and with term t_i.

The v_i in Equation (5) has three possible meaningful interpretations: (1) term t_i is a good discriminator when v_i is positive and large. This suggests that the introduction of term t_i decreases the document space density; (2) term t_i is a poor discriminator when v_i is negative and large. This suggests that the introduction of term t_i increases the document space density; (3) term t_i is an indifferent discriminator when v_i is near zero or equal to zero. This suggests that the introduction of term t_i has little (or no) impact on the document space density.

To make a visual judgment about a term discrimination value within the distance–angle-based visual environment, two visual displays of two document density spaces before and after that term assignment to the document collection must be produced, then the two visual displays are compared.

The following analysis is based on the distance-based similarity measure. That is, the similarity value between two documents in the collection is determined primarily by the distance between them in the document vector space. The nearer two documents are to one another, the more closely related they are.

The visual display of the document space density before term t_i assignment to the document collection is generated when the origin of the document vector space is selected as the minor reference point and the document centroid D_c is selected as the major reference point for the projection. In this case, the *Y*-axis value of a document in the visual space is defined as the distance from the document to the document centroid D_c, and the *X*-axis value of the document is the angle formed by the document and document centroid D_c against the origin in the document vector space. The document centroid is always projected onto the origin of the visual space in this projection mode. It is clear that the distribution of all documents along the *Y*-axis in the visual space reflects the document space density with term t_i.

The visual display of the document density space without term t_i is yielded when the origin of the document vector space and the document centroid D_{ci} [see Equation (4)] are assigned as the minor reference point and the major reference point for the projection, respectively. Note that term t_i should be removed from the document vector *V* [see Equation (1)] before the projection. In other words, *V* is modified as:

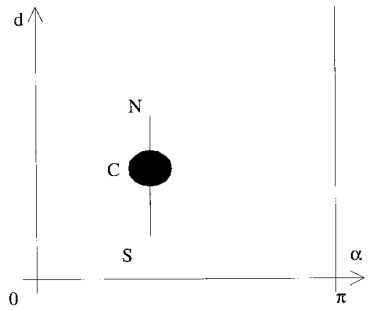

FIG. 4. Cluster movement along NS.

$$V_i = \begin{Bmatrix} t_{11} & t_{12} & \ldots t_{1i-1} & 0 & t_{1i+1} \ldots & t_{1n} \\ t_{21} & t_{22} & \ldots t_{2i-1} & 0 & t_{2i+1} & \cdots \\ \cdots & \cdots & & \cdots & & \cdots \\ t_{m1} & t_{m2} & \ldots t_{mi-1} & 0 & t_{mi+1} & t_{mn} \end{Bmatrix} \tag{6}$$

The distribution of all documents along the *Y*-axis in the visual space reflects the document space density without the term t_i.

After the two pictures are compared, the decision on the term discrimination capacity is made. Take, for example, the possible outcomes if term t_i is eliminated from the database and is then compared to the cluster distribution that contains t_i. If the document clusters (*C* in Fig. 4) in the visual space shift higher along the *Y*-axis (or *SN* in Fig. 4), this suggests that the space density has decreased. Therefore, term t_i is a poor discriminator for the documents in the collection. If the document clusters move down along the *Y*-axis (or *NS* in Fig. 4) the space density has increased, so the term t_i is a good discriminator. If the document clusters in the visual space remain largely the same, removal of the term t_i does not have a large impact on the space density and is, therefore, an indifferent discriminator.

To facilitate the visual judgment, we can change the color of the document icons depending on the direction of the shift. For instance, document icons are colored red when they shift up; the document icons are colored blue when they shift down in the visual space; or, the document icons keep their original color when they do not shift or shift little after term t_i is removed from the database. In this way, the colors of the document icons, the number of color-changed document icons, plus the positions of these document icons in the visual space give a clear overall picture of document density space changes.

An Application of Visual Term Discrimination Analysis

A full-text database from an *Associated Press (AP)* news file was utilized. The database consisted of 450 news reports dating from 1989. Forty-four distinct indexing terms were used to index the records.

Figure 5 shows the visual display of all documents with a specified term in the distance–angle-based visual space. In

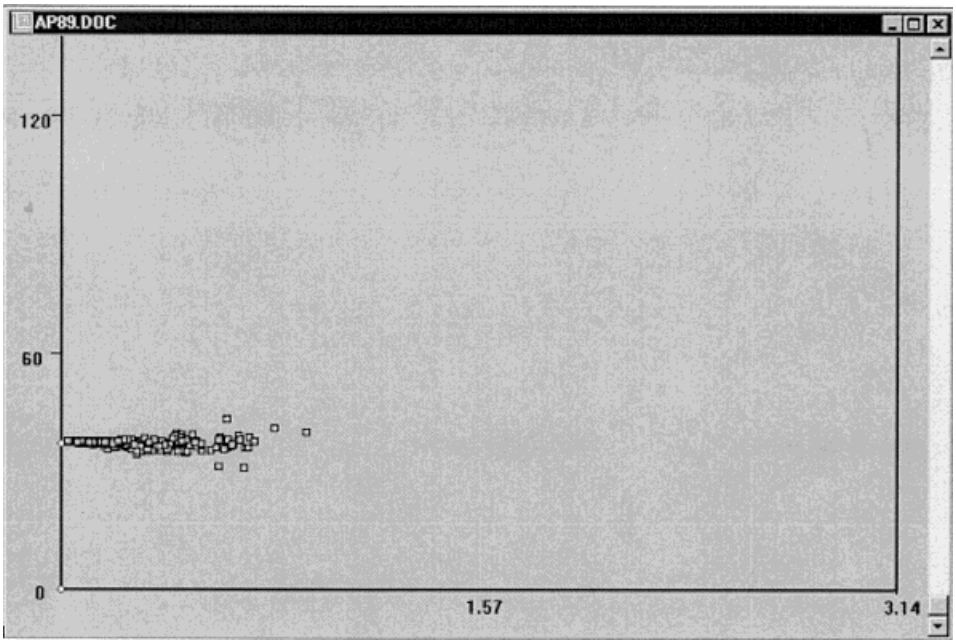

FIG. 5. Display of all documents with all terms against the distance-based similarity measure.

this case, no term was extracted from the database. This plot is used as the baseline against which other plots without a specified term t_i are compared.

Figure 6 displays the document distribution without a specified term t_i; in this case, the term is "Yeltsin." Observe that the document clusters move lower along the *Y*-axis. This implies that the term "Yeltsin" is a good discriminator for differentiating documents in this database. In other words, elimination of "Yeltsin" from the database has a visually significant impact on the document distribution.

In Figure 7, a document distribution is illustrated where the term that has been removed ("communist") is an indifferent discriminator. The document distribution within the visual space is almost the same as the baseline distribution.

Recall that the *DARE* visual space consists of two important parameters: visual distance and visual angle. We have demonstrated two examples using the distance-based measure [see Equation (7)]. Now let us also demonstrate how the discriminative capacity for the angle-based measure may be applied. To build the document density space using the angle-based similarity measure—the Cosine measure, for instance—the projection must change to the origin-based mode. According to the definition of the Cosine similarity measure, the origin of the document vector space

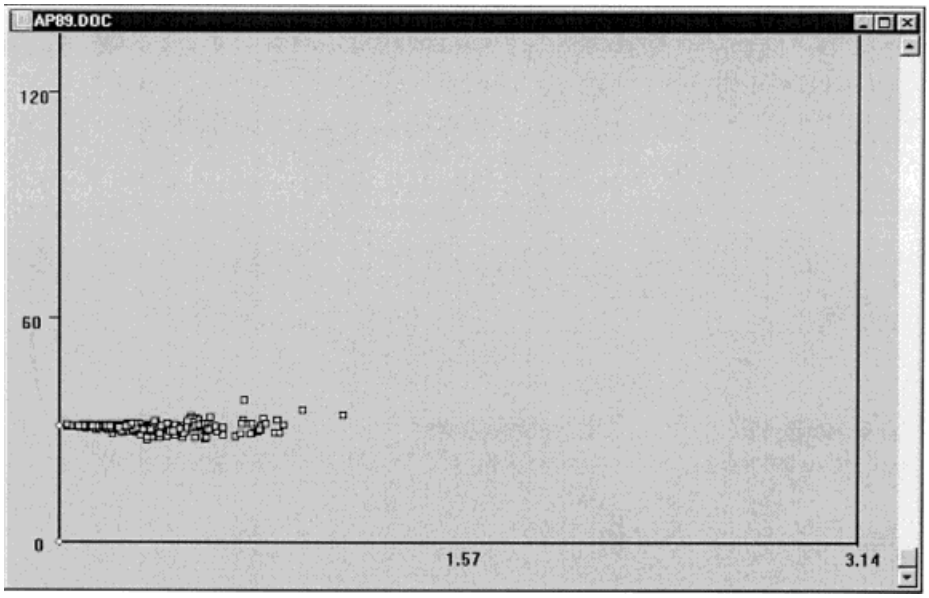

FIG. 6. Display of all documents without the specified term "Yeltsin" against the distance-based similarity measure.

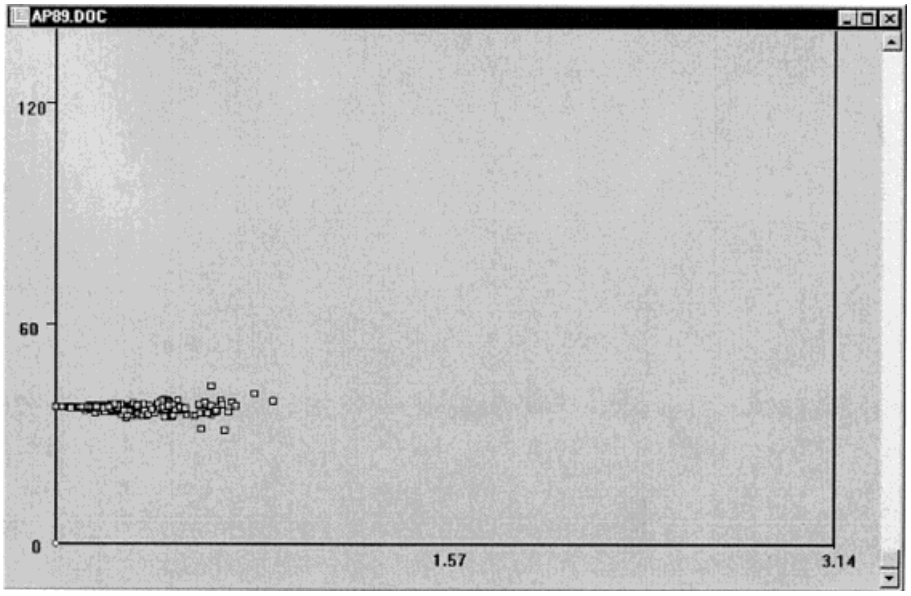

FIG. 7. Display of all documents without the specified term "communist" against the distance-based similarity measure.

must be assigned as the major reference point against which the visual angle of a document is measured. The minor reference point will be the document centroid. Because the reference points change, the document distribution in the visual space also changes. The way we judge the change of the document density spaces with and without a specified term will change accordingly. When document clusters in a document density distribution that exclude a specific term are shifted to the left, right, or remain unchanged within the visual document density space, the term is considered to be a good, poor, or indifferent discriminator, respectively. Only movement of document clusters along the *X*-axis in the visual space affects the term discriminative capacity for the angle-based similarity measure. The distances the document clusters move along the *X*-axis (*EW* or *WE* in Fig. 8) in the visual space reflect the extent to which the term serves as a discriminator to distinguish one term from another.

$$S(d) = \frac{1}{k^d} \tag{7}$$

In Equation (7), *d* is the distance between the object and the major reference point, *k* is a positive constant, and *S(d)* is its similarity value.

Figure 9 exhibits the visual document display with a specified term against the angle-based similarity measure. The document distribution is obviously different from that in Figure 5. The differences in the two pictures result from the exchange of the two reference points in the projection.

Note that the document density spaces under the distance-based similarity measure and the angle-based similarity measure should be measured and visualized in the two different projection modes. It is clear that term discrimination analysis under two different similarity measures needs four projections for all documents: two for the distance-based and two for angle-based measures. For efficiency, the document distribution changes for both measures may be integrated into a single plot. If the visual distance and visual angle of a document are redefined, the impacts of the two similarity measures on the term discrimination values can be illustrated in a same visual display. The *DARE* visual space does not currently support such a display, but it may be described conceptually here.

Toward this aim, the visual angle of a document is the same as its previous definition, and the visual distance is redefined as the distance from the document to the minor reference point rather than to the previous major reference point. In addition, the origin of the document vector space is assigned as the major reference point and the document centroid as the minor reference point in the projection. The impact of the document clusters in the newly proposed visual space on the term discriminative capacity may now be addressed.

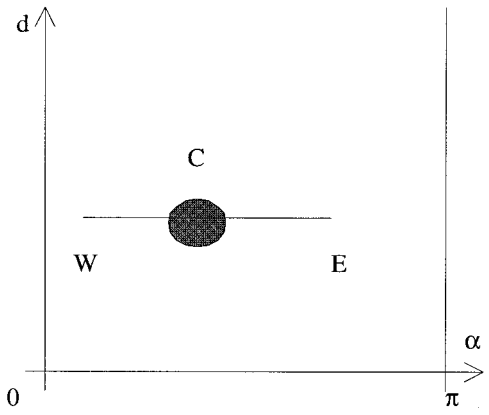

FIG. 8. Cluster movement along EW.

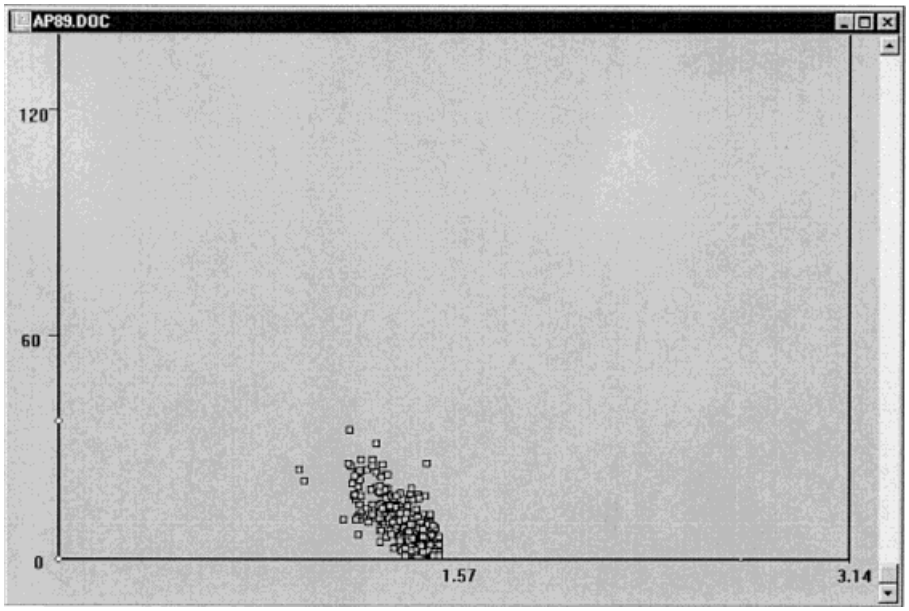

FIG. 9. Display of all documents with all terms against the angle-based similarity measure.

Possible movements in the document cluster based on differences in the distance and angle measures appear in Figure 10, where *C* is a document cluster and *di* ($i = 0, 1, 2, \ldots, 8$) refers to the movement direction. In the case of *d*0, the cluster remains unchanged after the removal of a term. The impact of changes in the measures on the term discrimination value appears in Table 1. The "+" and "−" symbols represent the positive and the negative effect, respectively, on the discriminatory impact of a term. For example, when the cluster moves in the direction *d*4, the specified term is considered to be a good discriminator for both the distance-based and the angle-based similarity measure.

Term Density Space Analysis

A term discriminative capacity should include two perspectives: to distinguish one document from other documents in the document density space, and to discern one indexing term from other terms in the term density space. Most research on term discrimination has only focused on the former. Term discrimination values for distinguishing one term from other terms can be used for term cluster analysis in a document collection.

A term density space can be built upon a term-document vector V' in Equation (8), representing the reversed vector of Equation (1).

$$V' = \begin{Bmatrix} t_{11} & t_{21} & \cdots & t_{m1} \\ t_{12} & t_{22} & \cdots & t_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ t_{1n} & t_{2n} & \cdots & t_{mn} \end{Bmatrix} \quad (8)$$

Two visual term density spaces, with and without a specified term, are created within the distance–angle-based visual environment so that the term discrimination value for distinguishing one term from others can be judged visually. The procedures for creating the two visual term density

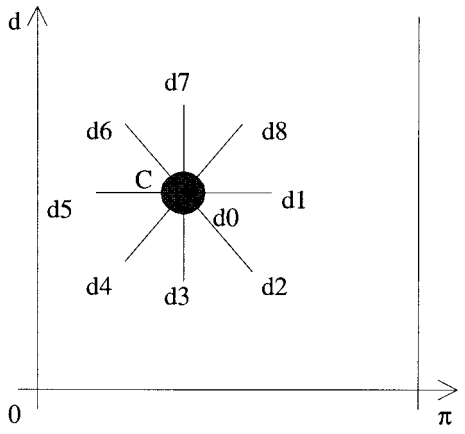

FIG. 10. Movement of a cluster in the modified visual space.

TABLE 1. Impact of the cluster movement on the discrimination in the modified visual space.

Directions	Impact of the distance-based measure on the discrimination	Impact of the angle-based measure on the discrimination
*d*1		−
*d*2	+	−
*d*3	+	
*d*4	+	+
*d*5		+
*d*6	−	+
*d*7	−	
*d*8	−	−
*d*0		

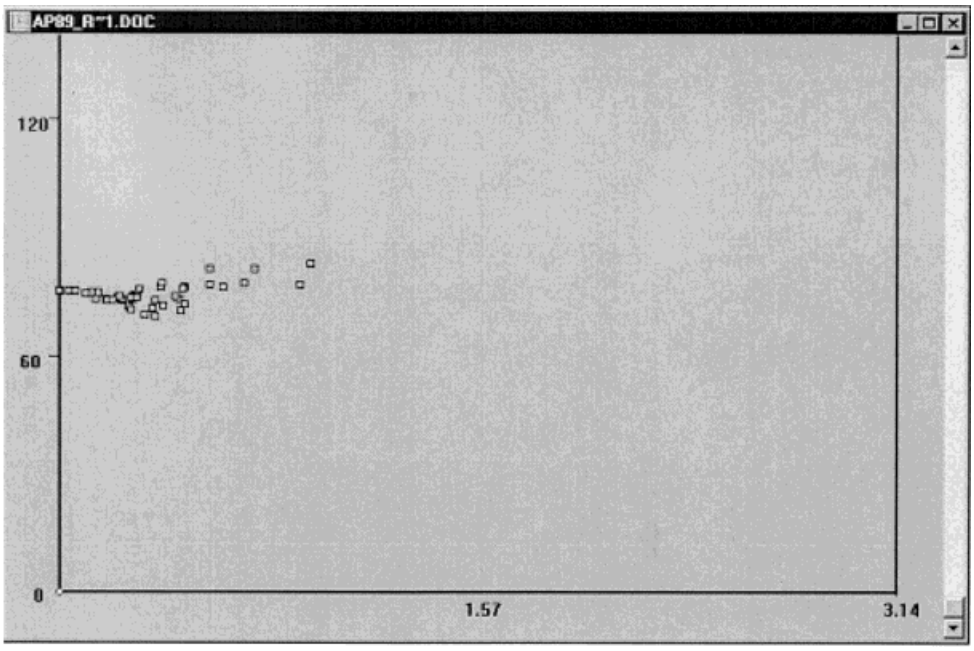

FIG. 11. Display of all documents with a specified term against the distance-based similarity measure.

spaces (with and without a term) are the same as those for the document density spaces. The only difference between them is that the visualized objects in the two visual spaces are different. The objects in the term density space are terms while the objects in the document density space are documents. The analysis can also be made against both the distance-based similarity measure and the angle-based similarity measure in the distance–angle-based visual space.

Figures 11 and 12 represent the visual term density displays with a specified term against the distance-based similarity and angle-based similarity measure, respectively.

Figure 13 displays the distribution of terms without the term "quake" against the distance-based similarity measure. It is concluded that the term "quake" is a good discriminator after the resulting distribution is compared with the standard display in Figure 11.

Unlike the document cluster movement in Figure 6, the term clusters in the visual space do not move downward evenly. The clusters near the *Y*-axis shift down more (that is, these clusters move closer to the centroid in the document collection, making a larger contribution to the term discriminative capacity), while clusters far from the *Y*-axis remain unchanged. This phenomenon can only be observed with the visual environment. Furthermore, from the plot, users can tell which terms in the density space are impacted and which are not by removal of a specified term. The traditional

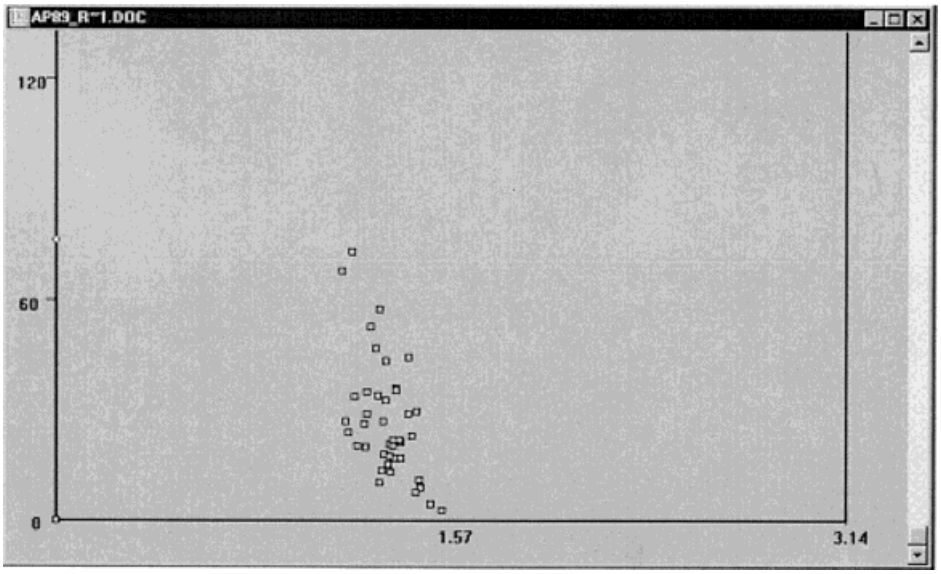

FIG. 12. Display of all documents with a specified term against the angle-based similarity measure.

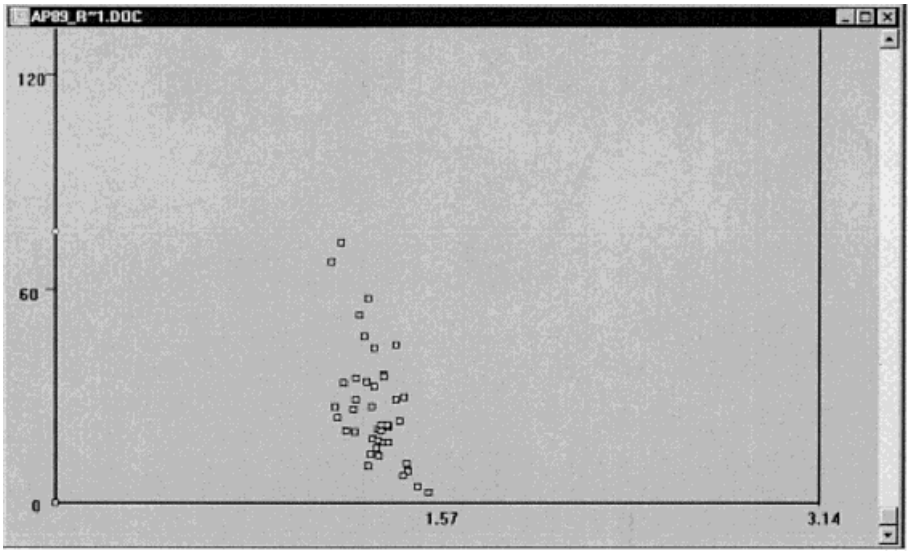

FIG. 13. Display of all documents without the term "quake" against the distance-based similarity measure.

approach for term discrimination analysis cannot provide users with this kind of information.

The overall results of this study for the *AP* database appear in Table 2, where *GD, PD,* and *ID* stand for good discriminator, poor discriminator, and indifferent discriminator, respectively.

The findings suggest that most of the terms in this dataset are indifferent discriminators. A good (poor) discriminator against one similarity measure within a document (term) density space does not necessarily correspond to a good (poor) discriminator against the other similarity measure within the same space. This is understandable because the similarity criteria for the distance-based measure are different from those of the angle-based measure, and sometimes they are not compatible. A good (poor) document discriminator against one similarity measure within a density space (for instance, the document density space) does not necessarily correspond to a good (poor) term discriminator against the same similarity measure within the other density space (the term density space). This indicates that the term density space has different characteristics from those of the document density space.

Discussion

The findings have applications for the development of more effective information retrieval systems and implications for both document indexing and system usage. Knowledge of term discrimination values can be useful in assigning weights to index terms in automatic indexing environments, where terms with the highest discrimination values receive the highest weights. Access to term discrimination values may also benefit searchers. The frequency of occurrence of index terms, commonly reported in information retrieval systems, does not provide a comprehensive picture of the potential usefulness of a specific term. By revealing the discriminatory characteristics of specific index terms, users will be able to make more informed decisions when selecting terms for query formulation.

The result of conventional term discrimination analysis is a single value. The discrimination value of a term does not show the impact of its discriminative capacity on an individual document or document cluster in the document collection. Therefore, a user is unaware of the impact of a given term on parts of the document collection because these differences are not revealed within a single value. In comparison to the conventional approach, the visual approach has the following added advantages:

(1) *The newly developed visual approach reveals characteristics of the document space that are normally hidden*. The document collection is no longer a "black box" for users (i.e., indexers and searchers). Document cluster movement before and after a specified term is pulled out from the document collection can be visualized in a two dimensional space. As indicated earlier, these changes in document position may be highlighted through the use of different colors for document icons to indicate the direction of the change. This will facilitate the identification of subtle movement of document clusters.

TABLE 2. Results of the term discrimination analysis.

	Distance-based similarity			Angle-based similarity		
	GD	PD	ID	GD	PD	ID
Term density space	2	2	40	0	0	44
Percentage	4.5%	4.5%	91%	0%	0%	100%
Document density space	1	0	43	4	0	40
Percentage	2.3%	0%	97.7%	9.1%	0%	90.9%

(2) *The impact of a term discrimination capacity on a specified document cluster in the document collection can be visually identified*. Suppose there are two terms with the same discrimination value in a document collection. The discriminative impacts of the two terms on the document collection may be significantly different. Discriminative capacity can impact all documents evenly or it can have a strong impact on only parts of document collection. These differences can be effectively demonstrated in the visual space. For instance, if all document icons move down (up) evenly along the *Y*-axis after a term is pulled out, it indicates that the discriminative capacity of the term impacts all documents evenly with respect to the distance-based similarity measure. If only parts of document clusters move along the *Y*-axis and the rest of the documents are unaffected, this implies that the term discriminative capacity impacts only the vertically moving document clusters with respect to the distance-based similarity measure. From this observation it can be concluded that if documents from the moving document clusters are indexed with a specific term in which this behavior is observed, these documents should be assigned different discrimination values (based on the degree and the direction of the document cluster movement) from those that do not move. However, in the conventional approach, all documents are treated equally. That is, each document is assigned the same discrimination value regardless of whether or not it makes a contribution to the discriminative capacity. Another example also demonstrates the weakness of the traditional approach. Due to the removal of a term from the document collection, some documents move toward the centroid and some move away from the centroid. The combination of the document movements may result in a discrimination value of zero. The effects of opposite document movement directions cancel each other out. On the other hand, it is clear that if there is no document cluster movement without considering the term, then its discrimination value is zero as well. Within the visual environment, the two situations are clearly distinguished. Therefore, the term is recognized as a good/poor discriminator based on the document cluster movement in the opposite directions along the *Y*-axis respectively. Unfortunately, the term, which is recognized as an indifferent discriminator for all documents in the conventional approach, cannot discriminate these two situations simply because they have the same discrimination value (zero). Within the visual environment, users can also examine the discriminative impact of a term on a specified document cluster in the document collection by displaying only the movement of this document cluster without that term. It gives more flexibility for term discrimination analysis. It is apparent that the traditional approach cannot make such a "local discrimination analysis" due to the fact that document clusters cannot be identified directly in a high dimensional document space. These features illustrate the uniqueness of the visual approach.

(3) *It is widely recognized that the distance-based similarity measure and the angle-based similarity measure have distinct characteristics*. In other words, each has its own strength for information retrieval. The conventional term discrimination approach provides users with term discrimination values against only one similarity measure. In this case, the visual tool has the potential to allow users to make the decision about the term discriminative capacity against the two similarity measures.

To effectively display a large number of document icons in a limited screen area is a common and universal problem for all visual tools, including this visual term discrimination analysis tool. There are several possible solutions to this problem. One solution is to use a zoom mechanism to distinguish overcrowded document icons. In our case, to reduce the number of projected documents in the visual space, the system can display the documents whose positions change after the term is pulled out from the document collection, and filter out the documents that do not move.

The visual term discrimination analysis interface is completely compatible with the visual information retrieval interface. It can be easily integrated into the visual information retrieval interface (*DARE*) as an additional feature.

Although primarily intended for indexing IR system contents, the application of visual term discrimination analysis techniques may extend to the search environment, where visualization of term clusters may be useful in identifying additional relevant terms during query formulation. Within the same environment, if searchers submit queries instead of the term centroids, the system would generate a document distribution based on the queries. Users can then search relevant documents via different information retrieval models within the visual environment. They may also browse documents or analyze document clusters.

Conclusion

DARE, a visual information retrieval tool, has been employed to visualize the document density space with and without a specified term. The document distribution changes before and after a term assignment provide users with overall pictures for the term discrimination capacity. Using different document projection modes (the origin-based or nonorigin-based), the discrimination values of a term against the different similarity measures (the distance-based similarity measure or angle-based similarity measure) can be analyzed visually within the distance–angle-based visual information environment. To date, no visual information tool has been designed for term discrimination analysis. Term discrimination analysis research has concentrated on its application for algorithm efficiency improvement. The introduction of visualization techniques to the area of term discrimination analysis opens a new and unexplored area of study.

A possible improvement for the visualization of term discrimination analysis has been discussed, where the visual distance of a projected document in the document vector space is redefined as the distance from the measured docu-

ment to the minor reference point. The benefit of this modification is that term discrimination analyses against both the distance-based similarity measure and the angle-based similarity measure may be visualized in a single visual space. In addition, the combination offers even richer information for users.

The authors have demonstrated that the term density space for a document collection can also be visually demonstrated within *DARE*'s visual space. Following the same procedures as in the document density space analysis, the term discrimination capacity analysis can also be visualized. A term discrimination capacity analysis from both the document density space and the term density space perspectives provides a more objective and complete understanding of the nature of the interrelationships among terms and documents.

The study shows that most of the terms in the employed dataset are indifferent discriminators. A good (poor) discriminator against one similarity measure within a density space does not necessarily correspond to a good (poor) discriminator against the other similarity measure within the same density space. Similarly, a good (poor) document discriminator against one similarity measure within a density space does not necessarily correspond to a good (poor) term discriminator against the same similarity measure within the other density space. Within the visual document (term) density space, users cannot only tell whether a specified term is a good, poor, or indifferent discriminator, but also can tell which documents (terms) are impacted by the term, which are not impacted, and the degree to which the documents (terms) are impacted. Visualization is a convenient, effective, and efficient way for the term discrimination analysis.

The authors are cautious about drawing general conclusions regarding the distribution of discrimination values within databases. The primary objective of this study was to determine whether or not the visualization approach is technically feasible for term discrimination analysis. The test collection used consisted of a relatively small number of records of news-related full-text documents with a small number of index terms. The distribution of discrimination values may well vary with the size and coverage of the database as well as the indexing exhaustivity, which may impact the ultimate utility of these methods. Further testing on a broad range of collections is needed. Future research will investigate the application of the methods discussed using larger and more diverse data sets. The next phase of this research will investigate document collections containing a larger number of index terms with different term distribution and exhaustivity characteristics to determine the impact of these factors on term discriminative capacity. The visualization methods for cluster movement will also be integrated into the *DARE* environment. Additional methods for visualization and metrics for the summarization of term discrimination value distributions will also be developed. Following more rigorous investigation of visualization of term discriminative capacity, a user study will be carried out to gauge indexer and searcher affective response to the use of visualization methods for this purpose.

Acknowledgments

The authors would like to acknowledge the anonymous referees for their helpful comments.

References

Biru, T., El-Hamdouchi, A., & Rees, R.S. (1989). Inclusion of relevance information in the term discrimination model. Journal of Documentation, 45, 85–109.

Borner, K. (2000). Extracting and visualizing semantic structures in retrieval results for browsing. In Proceedings of the fifth ACM conference on digital libraries (pp. 234–235). San Antonio: ACM.

Brooks, M., & Campbell, J. (1999). Interactive graphic queries for bibliographic search. Journal of the American Society for Information Science, 50(9), 814–825.

Chang, J.W., Lee, J.H., & Lee, Y.J. (1989). Multikey access methods based on term discrimination and signature clustering. In Proceedings of the 12th annual international ACM SIGIR conference on research and development in information retrieval (pp. 176–185). Cambridge, MA: ACM.

El-Hamdouchi, A., & Willett, P. (1988). An improved algorithm for the calculation of exact term discrimination values. Information Processing and Management, 24(1), 17–22.

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.

Fowler, R.H., Fowler, W.A.L., Wilson, B.A. (1991). Integrating query, thesaurus, and documents through a common visual representation. In Proceedings of the fourteenth annual international ACM SIGIR conference on research and development in information retrieval (pp 142–151). Chicago: ACM.

Furnas, G.W., & Rauch, S.J. (1998). Considerations for information environments and the NaviQue workspace. In Proceedings of the third ACM conference on digital libraries (pp. 79–88). Pittsburgh: 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.

Heath, L., Hix, D., Novell, L.T., Wake, W.C., Averboch, G.A., Labow, E., Guyer, S.A., Brueni, D.J., France, R.K., Dalai, K., & Fox, E.A. (1995). Envision: A user-centered data base of computer science literature. Communications of the ACM, 38(4), 52–53.

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

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. (1997). Information storage and retrieval. New York: Wiley Computer Pub.

Lin, X. (1993). A self-organizing semantic map for information retrieval. Ph.D. Dissertation, University of Maryland at College Park.

Michard, A. (1982). Graphical presentation of Boolean expressions in a database query language: Design notes and an ergonomic evaluation. Behavior and Information Technology, 1(3), 279–288.

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.

van Rijsbergen, C.J. (1979). Information retrieval. Boston: Butterworths.

Willett, P. (1985). An algorithm for the calculation of exact term discrimination values. Information Processing and Management, 21(3), 225–232.

Wise, J.A. (1999). The ecological approach to textual visualization. Journal of the American Society for Information Science, 50(13), 1224–1233.

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. (1999). Visual information retrieval environments. Ph.D. Dissertation, School of Information Sciences, University of Pittsburgh.

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.