PERGAMON

Information Processing and Management 37 (2001) 279–294

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Developing a new similarity measure from two different perspectives

Jin Zhang [a,*], Edie M. Rasmussen [b]

[a] *School of Library and Information Science, University of Wisconsin–Milwaukee, 2400 East Hartford Avenue, Enderis Hall, Milwaukee, WI 53211, USA*

[b] *School of Information Sciences, University of Pittsburgh, 135 North Bellefield Avenue, Pittsburgh, PA 15260, USA*

Received 21 July 1999; accepted 9 May 2000

Abstract

In this paper two distinct similarity measures in a document vector space, the distance-based and angle-based similarity measures, are compared, and a newly developed similarity measure based upon both the distance and angle strengths of two compared objects is presented. The concept of the iso-extent contour, which facilitates the understanding of the nature of the newly developed similarity measure, is introduced. The three different similarity measures are compared and the properties of the newly developed similarity measure are addressed.

Keywords: New similarity measure; Angle-based similarity measure; Distance-based similarity measure

1. Introduction

The similarity measure is an important concept in information retrieval. It is a measure used to compare two objects and to determine whether they are related to the same topics. The two objects can be a reference point and a document or two documents in a document collection. Different information retrieval systems have different approaches to measuring the similarity between two objects. In a vector-based information retrieval system, spatial distance and the relative direction of two objects are the measures most used to judge similarity.

Similarity is associated with relevance but they are different in some respects. For instance, similarity between two objects is an objective measure while their relevance may be subjective. A similarity measure is employed to determine whether a document is likely to be relevant to a

* Corresponding author.
E-mail address: jzhang@uwm.edu (J. Zhang).

0306-4573/01/$ - see front matter
PII: S0306-4573(00)00027-3

specified reference point or query, and compared documents are retrieved if the similarity value is larger than the threshold specified for a search. It may also be applied to rank retrieved documents for users, allowing them to effectively decrease or increase the number of a retrieved document set. Similarity measures are used in other situations as well, such as to automatically classify documents based on similarities among them in a document collection, and to create document clusters. They are also used in building visual retrieval environments such as *GIUDO* (Nuchprayyoon, 1996), *VIBE* (Olsen et al., 1993; Olsen & Korfhage, 1994), and *DARE* (Zhang & Korfhage, 1999). Construction of a rational, reliable and practical similarity measure is a fundamental and substantial research topic in the field of information retrieval.

According to McGill, Koll and Noreault (1979), there are more than 60 different similarity measures, and the number keeps increasing. These include the inner product similarity measure (Falkowski, 1998), Dice coefficient, cosine coefficient, Jaccard coefficient, overlap coefficient (Salton, 1968, 1989; Meadow, 1992; Frakes & Baeza-Yates, 1992; Korfhage, 1997; Rousseau, 1998), the spreading activation similarity measure (Jones & Furnes, 1987), and some probability-based similarity measures (Croft & Harper, 1979; Van Rijsbergen, 1979; Radecki, 1982; Kwok, 1985; Robertson & Sparck Jones, 1976; Robertson & Walker, 1997). Some measures use query feedback information (Bartell, Cottrell & Belew, 1998), citation information (Travison, 1987; Cronin, 1994), or semantic closeness (Tudhope & Taylor, 1997) to determine the similarity of compared objects. Other research on this topic includes Radecki (1985), Griffiths et al. (1986), Ellis et al. (1993), Tudhope & Taylor (1996), and Fricke (1997). The most popular measures are the distance-based similarity measure and the angle-based cosine measure. Zhang has developed a distance-angle integrated similarity measure (Zhang & Korfhage, 1999) which uses an angle strength modifier to modify a traditional distance-based similarity measure.

Research in the field has focused on developing different kinds of similarity measures based on different information retrieval systems. Very little work (Hamers et al., 1989) has been done on a comparison between the distance-based and the angle-based similarity measures in a vector-based information retrieval system and the development of new similarity measures based on the strengths of the two major similarity measures.

The aims of this paper are to present a new similarity measure which takes advantage of both the distance-based and angle-based similarity measures and overcomes their inherent weaknesses, to address the rationale for the measure from an information retrieval perspective, to compare it with the distance-based and angle-based similarity measures, and to investigate properties of the new similarity measure.

2. Proposed similarity measure and its rationale

Before introducing the newly developed similarity measure, we will analyze the two distinct similarity measures, the distance-based similarity measure and the angle-based similarity measure, from an information retrieval perspective and elicit useful information for the development of the

new similarity measure. Since the development of the new similarity measure is based on the two major similarity measures, the analyses demonstrate the motivation for developing the new similarity measure and provide insight into its characteristics.

The distance-based similarity measure follows the principle that a document in a vector document space close to a reference point is likely to be relevant to it. The farther from the reference point the document is, the less relevant or similar it is and vice versa. In the extreme case the document overlaps with the reference point and they are exactly similar. The principle is easily understood, an intuitive and practical approach. The distance-based similarity measure judges the similarity between two objects only from the geographic distance between two objects regardless of the direction of the two objects vis-à-vis the origin of the document vector space.

However, the direction of two objects vis-à-vis the origin definitely has impact on the similarity between them, which is the basis of the angle-based similarity measure. The characteristics of the angle-based similarity measure are discussed in detail below.

There may be many different ways to explain the Cosine similarity measure. The "angle" is one of them. By its mathematical definition, the Cosine similarity measure is an angle-based or direction-based similarity measure. Direction is a fundamental characteristic in a vector space. The origin in vector space as one base point plays an important role in determining the direction of a document. In the Cosine measure it is the direction of two documents that is used to judge the relevance between them. In other words, the origin directly affects the directions of measured documents and then affects the relevance judgement between them. In addition, the length of documents is determined by weighting in the vector space which is defined against the origin. The origin is a base point for assigning weights to a document. Without the origin, weights of a document are meaningless, the length of a document is meaningless.

Suppose R is a reference point, Di and Dj are two documents in a document vector space:

$R(x_{k1}, x_{k2}, \ldots, x_{kn})$;

$Di(x_{i1}, x_{i2}, \ldots, x_{in})$; and

$Dj(x_{j1}, x_{j2}, \ldots, x_{jn})$.

Where $x_{jr} = c * x_{ir}$, $(r = 1, \ldots, n)$, c is a constant and n is the dimensionality of the document vector space.

It is clear that documents Di and Dj lie in the same direction vis-à-vis the origin of the document vector space due to the same proportional weights of each pair of keywords.

The documents Di and Dj are exactly similar to the reference point R if a cosine function is applied as an angle-based similarity measure.

$$\cos(R,\ Di) = \frac{\sum_{r=1}^{n} x_{kr} * x_{ir}}{\left(\sum_{r=1}^{n} x_{kr}^2\right)^{1/2} * \left(\sum_{r=1}^{n} x_{ir}^2\right)^{1/2}} \tag{1}$$

$$\cos(R,\ Dj) = \frac{\sum_{r=1}^{n} x_{kr} * x_{jr}}{\left(\sum_{r=1}^{n} x_{kr}^2\right)^{1/2} * \left(\sum_{r=1}^{n} x_{jr}^2\right)^{1/2}}$$

$$\cos(R,\ Dj) = \frac{\sum_{r=1}^{n} k_{kr} * c * x_{ir}}{\left(\sum_{r=1}^{n} x_{kr}^2\right)^{1/2} * \left(\sum_{r=1}^{n} c^2 * x_{ir}^2\right)^{1/2}} \tag{2}$$

$$\cos(R,\ Dj) = \frac{\sum_{r=1}^{n} x_{kr} * x_{ir}}{\left(\sum_{r=1}^{n} x_{kr}^2\right)^{1/2} * \left(\sum_{r=1}^{n} x_{ir}^2\right)^{1/2}}$$

$$\cos(R,\ Di) = \cos(R,\ Dj)$$

Eq. (2) shows that when documents have the same keyword distribution and proportional weights for each keyword, they are exactly similar in terms of the angle-based similarity measure. If keyword weights within a document are determined primarily by the keyword occurrences within that document, documents with different lengths but the same topics determined by the same keyword distribution are easily identified by using the angle-based similarity measure.

Suppose that D*i* is fixed and *Dj* is movable. The distance *d* from the document *Di* to document *Dj* is defined as:

$$d = \left(\sum_{r=1}^{n} (x_{ir} - x_{jr})^2\right)^{1/2}$$

$$d = \left(\sum_{r=1}^{n} (x_{ir} - c * x_{ir})^2\right)^{1/2} \tag{3}$$

$$d = |1 - c| * \left(\sum_{r=1}^{n} (x_{ir})^2\right)^{1/2}$$

Eq. (3) indicates that the distance between the two documents depends only on the proportion parameter *c*. Eqs. (2) and (3) suggest that even when the documents *Di* and *Dj* are far apart (a large proportion parameter *c*), the similarity between them remains the same. In other words, the similarity between the documents *Di* and *Dj* is not related to the proportion parameter *c* which determines the distance between the two documents.

However, when the distance between two objects with the same direction vis-à-vis the origin of a document vector space is great enough, the validity of the angle-based similarity between the two objects is in question. It is suggested that the similarity between two objects should be judged not only by the topics they address but also by the extent of their coverage of the same topics. For instance, there is a significant difference between an article in a professional journal and the same topic covered in an article in a newspaper. Neither of the two perspectives (the topics themselves

and the extent of the coverage) should be ignored in judgement of the similarity between two objects. Otherwise the accuracy and validity of a similarity measure could be challenged. This inherent weakness of the angle-based similarity measure comes from a lack of consideration of the contribution of the distance parameter when comparing the similarity of two objects.

To illustrate this point we introduce the concept of the iso-extent contour. The iso-extent contour refers to the contour in a document vector space on which all documents are similar in terms of the extent to which they address topics. The topics they address are not necessarily the same. The distance from the origin of the document vector space to any of the points on an iso-extent contour is the same, indicating that the iso-extent contour is a circle whose central point is the origin of the document vector space. If weights of keywords of a document are positive, the iso-extent contour is located only in the first quadrant of the document vector space (see Fig. 1). There are countless iso-extent contours in a document vector space.

The distance from a document to the origin of a document vector space can be used to measure the extent to which that document covers its topics. Therefore, the difference in distance of two objects to the origin can be employed to measure the extent of the difference between them. It is obvious that the farther apart two documents *Di* and *Dj* are, the larger the difference in the extent of their coverage of the same topics. The larger this difference, the less similar to each other they are in terms of the extent of their coverage of the topics.

The concept of the iso-extent contour helps in understanding the nature of the angle-based similarity measure which shapes its inherent strength and weakness.

In the application of the angle-based similarity measure, assume that there is a threshold α for a search. Documents within the angle area determined by the threshold α are regarded as retrieved documents (see Fig. 2). It is clear that the retrieved area is unlimited.

Suppose that *D* is a document on one line bounding the angle area, and *C*1 and *C*2 are two points on the other line bounding it. According to the traditional angle-based similarity measure, the similarity between *D* and *C*1 is the same as that between *D* and *C*2 because *C*1 and *C*2 have

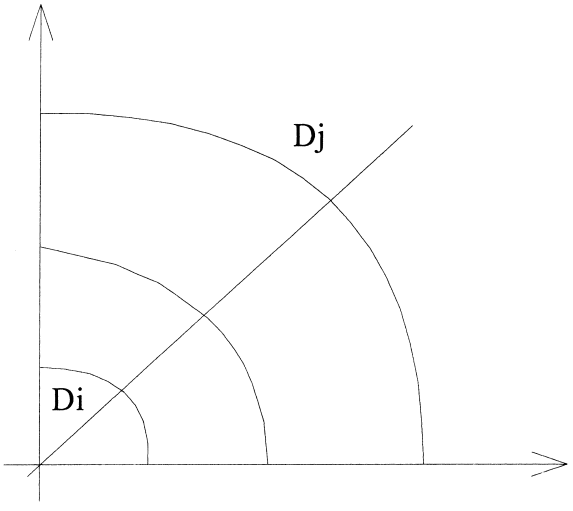

Fig. 1. Iso-extent contour.

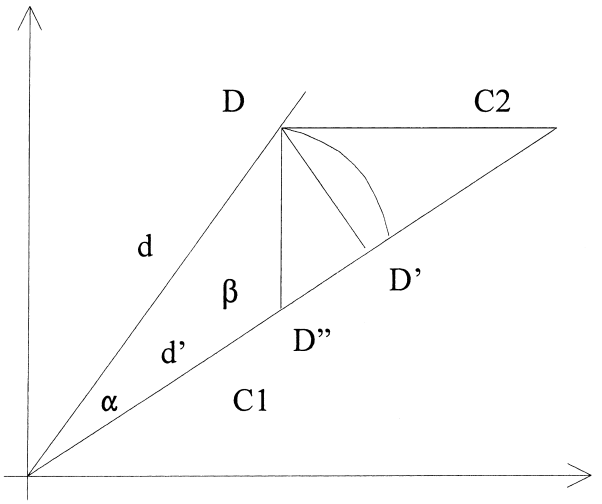

Fig. 2. Display of the angle-based similarity measure.

the same direction vis-à-vis the origin of the document vector space no matter how far apart they are. Generally speaking, any two points on the line forming the angle are exactly similar to D with respect to the angle-based similarity measure.

In order to develop a new reasonable and reliable similarity measure, we must solve this problem by considering the contribution of distance from the origin to the similarity of the two objects.

The point D' is the intersection between the line $C1C2$ and the iso-extent contour of the point D. Suppose that $C1$ is point being compared to D', d is the distance from the origin of the document vector space to D and d' is the distance from the origin to $C1$. The new similarity measure is defined in Eq. (4) or (5).

$$S = a^{|d-d'|} * \cos(\alpha) \tag{4}$$

$$S = a^{|(\Sigma_{r=1}^{n}(x_{ir})^{1/2})^2-(\Sigma_{r=1}^{n}(x_{jr})^{1/2})^2|} * \frac{\sum_{r=1}^{n} x_{ir} * x_{jr}}{\left(\sum_{r=1}^{n} x_{ir}^2\right)^{1/2} * \left(\sum_{r=1}^{n} x_{jr}^2\right)^{1/2}} \tag{5}$$

Suppose:

$D(x_{i1}, x_{i2}, \ldots, x_{in})$; and

$C1(x_{j1}, x_{j2}, \ldots, x_{jn})$.

Where a is a constant used to control the extent to which the distance strength impacts on the similarity measure, n $(r = 1, \ldots, n)$ is the dimensionality of the document vector space. The constant a is equal to or less than one. Since the similarity value is always positive, the constant is required to be positive $(0 < a \leqslant 1)$. When a is equal to one, there is no impact of the distance parameter on the similarity measure. The larger the constant a, the stronger the impact of the distance parameter on the similarity measure, and vice versa. Because $0 < a^{|d-d'|} \leqslant 1$ $(0 < a \leqslant 1)$ and $0 \leqslant \cos(\alpha) \leqslant 1$ $(0 \leqslant \alpha \leqslant \pi/2)$, we have $0 \leqslant S \leqslant 1$.

Here $|d-d'|$ is regarded as a measure of the extent of the difference between the two objects, and $a^{|d-d'|}$ is a modifier of the distance parameter. It is used to modify the angle-based similarity measure. When $C1$ is far from D', the relatively small value of the modifier $a^{|d-d'|}$ makes the similarity value decrease. When $C1$ moves along the line $C1O$ toward D', the similarity value increases due to the increasing value of the modifier $a^{|d-d'|}$. When $C1$ overlaps with D', the value of the modifier is equal to one. This suggests that the object being compared is on the iso-extent contour of D, and the impact of the distance parameter is zero.

The role of $a^{(d1-d2)}$ in the newly developed similarity measure is to measure not the distance strength between the measured documents but the extent difference of their coverage of the topics which is dependent on the distances between the origin and the measured documents $d1$ and $d2$. It is designed to modify the traditional angle-based similarity measure from the extent perspective.

Consider another situation when D moves along its iso-extent contour toward D'. In this case, the angle α determined by D and $C1$ vis-à-vis the origin of the document vector space changes while the modifier $a^{|d-d'|}$ remains the same. This implies that the impact of the distance strength on the similarity measure is unchanged while the change of the similarity value is caused only by the change of the angle α.

When D moves along the line generated by D and $C1$, the situation is more complicated. The similarity value of this new similarity measure is affected by both the changing angle α and the modifier $a^{|d-d'|}$. In other words both the distance and the direction affect the similarity measure. The direction of the movement of D determines the extent to which the angle α and the modifier $a^{|d-d'|}$ impact the similarity measure.

Note that in some information retrieval systems, document lengths are normalized to the range 0 to 1. After normalization, the distance difference between two documents in the document vector space may not be as large as that between the two documents in the un-normalized document vector space. The maximum distance difference between two documents is one in this case. The phenomenon suggests that the extent difference between two documents in the document space is not as great as that in an un-normalized document space. However, normalization does not totally eliminate the distance difference between two documents. It simply changes the scale of the system, and basic document distance relationships in the document vector are still preserved. In addition, the impact of the difference in content of the two documents on the newly developed similarity measure can be adjusted by manipulating the parameter a if the approach is applied in a normalized document space.

3. Property analyses of the new similarity measure

3.1. Comparisons among the distance-based, angle-based, and new similarity measures

To illustrate the different performance of the three different similarity measures, we select a point D in a document vector space, to monitor the similarity value between D and $C1$ using each individual similarity measure, and to compare their similarities against the fixed $C1$. This will demonstrate the differences between the three similarity measures and show how the newly developed similarity measure compensates for the weaknesses of the other two.

First, the three similarity measures are compared and discussed when D moves from the origin of a document vector space alone the line OD. In this case the threshold angle α of a search remains the same and the iso-extent contour of the moving point D changes (see Fig. 2). The angle β is formed by the lines $DC1$ and $OC1$.

Observe that from Fig. 2 we can derive the following important equations

$$d * \sin(\alpha) = DC1 * \sin(\beta) \tag{6}$$

$$d * \cos(\alpha) - DC1 * \cos(\beta) = d' \tag{7}$$

From Eqs. (6) and (7):

$$\begin{aligned} &d^2 * (\sin(\alpha))^2 = DC1^2 * (\sin(\beta))^2 \\ &(d * \cos(\alpha) - d')^2 = DC1^2 * (\cos(\beta))^2 \\ &DC1^2 = d^2 - 2d * d' * \cos(\alpha) + d'^2 \\ &DC1 = (d^2 - 2d * d' * \cos(\alpha) + d'^2)^{1/2} \end{aligned} \tag{8}$$

$$DC1 = -(d^2 - 2d * d' * \cos(\alpha) + d'^2)^{1/2} \tag{9}$$

It is obvious that Eq. (9) is not valid because $DC1$ should be positive. Therefore, Eq. (8) is kept and Eq. (9) is discarded.

In this case, the distance-based similarity measure is defined in Eq. (10),

$$\begin{aligned} &s = a^{DC1} \\ &s = a^{(d^2 - 2d * d' * \cos(\alpha) + d'^2)^{1/2}} \end{aligned} \tag{10}$$

where α and d' are fixed during the movement and d is a variable.

In Fig. 3, $S1(d)$, $S2(d)$ and $S3(d)$ represent the newly developed similarity measure [Eq. (4)], the angle-based similarity measure ($\cos(\alpha)$), and the distance-based similarity measure [Eq. (10)], respectively. In Fig. 3, $d' = 10$, $\alpha = \pi/4$, $a = 0.99$. The X- and Y-axes are d and similarity, respectively.

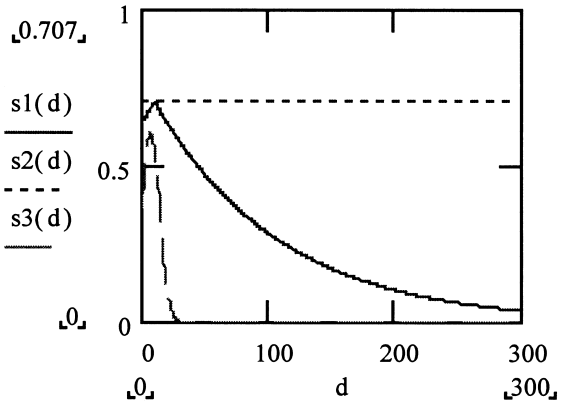

Fig. 3. Comparisons of the similarity measures (I).

The plot of the angle-based similarity measure is a horizontal line due to the un-changing angle α during the movement. The curve of the newly developed similarity measure shows that it reaches its maximum value when $C1$ overlaps with the iso-extent contour of the moving point D. At this point the two objects being compared share the same extent of coverage. The curve gradually approaches zero when d increases. The curve of the distance-based similarity measure is like that of the newly developed similarity measure. It also has a maximum value. One of the differences is that the curve reaches its maximum value a little earlier than that of the newly developed similarity measure. The distance-based similarity measure and the newly developed similarity measure reach their maximum values in D'' and D', respectively (see Figs. 2 and 3). The line DD'' is perpendicular to the line $C1O$. D'' is the nearest point on the line $C1O$ to D. The average similarity value of the distance-based similarity measure is smaller than that of the newly developed similarity measure in this case.

Consider the comparisons among the three similarity measures when the moving point passes along the line $DC1$. Fig. 2 demonstrates that when the moving point passes along $DC1$, the angle $\angle DC1O$ is fixed and the angle α changes. In order to compare and analyze the newly developed similarity measure, we must determine the relationship between the variables d and α.

From Eqs. (6) and (7):

$$\begin{aligned} & d * \cos(\alpha) - \frac{d * \sin(\alpha)}{\sin(\beta)} * \cos(\beta) = d' \\ & \sin(\beta - \alpha) = \frac{d' * \sin(\beta)}{d} \\ & d = \frac{d' * \sin(\beta)}{\sin(\beta - \alpha)} \end{aligned} \tag{11}$$

According to Eqs. (4) and (11):

$$s = a^{\frac{d' * \sin(\beta)}{\sin(\beta - \alpha)} - d' * \cos(\alpha)} \tag{12}$$

where the angle β and d' are fixed during the movement, the definitions of d, d' and α are the same as the previous definitions.

Suppose:

$D(x_{i1}, x_{i2}, \ldots, x_{in})$;
$C1(x_{j1}, x_{j2}, \ldots, x_{jn})$; and
$C2(x_{k1}, x_{k2}, \ldots, x_{kn})$.

$$\beta = \frac{\sum_{r=1}^{n} (x_{ir} - x_{jr}) * (x_{kr} - x_{jr})}{\left(\sum_{r=1}^{n} (x_{ir} - x_{jr})^2 \right)^{1/2} * \left(\sum_{r=1}^{n} (x_{kr} - x_{jr})^2 \right)^{1/2}} \tag{13}$$

Since the variable is the angle α rather than d in this case, the distance-based similarity measure should be expressed in the form of α.

From Eqs. (6) and (7):

$$DC1 = \frac{d * \sin(\alpha)}{\sin(\beta)}$$

From Eq. (11):

$$DC1 = \frac{d' * \sin(\alpha)}{\sin(\beta - \alpha)} \tag{14}$$

$$s = a^{DC1}$$

$$s = a\frac{d' * \sin(\alpha)}{\sin(\beta - \alpha)} \tag{15}$$

In Fig. 4, $S1(\alpha)$, $S2(\alpha)$ and $S3(\alpha)$ stand for the newly developed similarity measure [Eq. (12)], the angle-based similarity measure ($\cos(\alpha)$), and the distance-based similarity measure [Eq. (15)], respectively. In Fig. 4, $d' = 10$, $\beta = 3\pi/5$, $a = 0.9$. The X- and Y-axes are angle α and similarity, respectively.

Last, let us discuss the comparisons of the three similarity measures when the moving point D moves along its iso-extent contour. In this case d and d' are fixed while angle α is the variable. Therefore, Eq. (10) should be used to yield the curve of the distance-based similarity measure.

In Fig. 5, $S1(\alpha)$, $S2(\alpha)$ and $S3(\alpha)$ stand for the newly developed similarity measure [Eq. (4)], the angle-based similarity measure ($\cos(\alpha)$), and the distance-based similarity measure [Eq. (10)], respectively. In Fig. 5, $d = 20$, $d' = 10$, and $a = 0.99$. The X- and Y-axes are angle α and similarity, respectively.

In this case the curve of the newly developed similarity measure is much like that of the angle-based similarity measure.

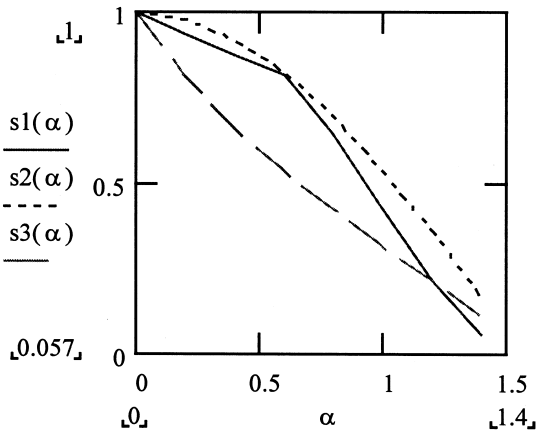

Fig. 4. Comparisons of the similarity measures (II).

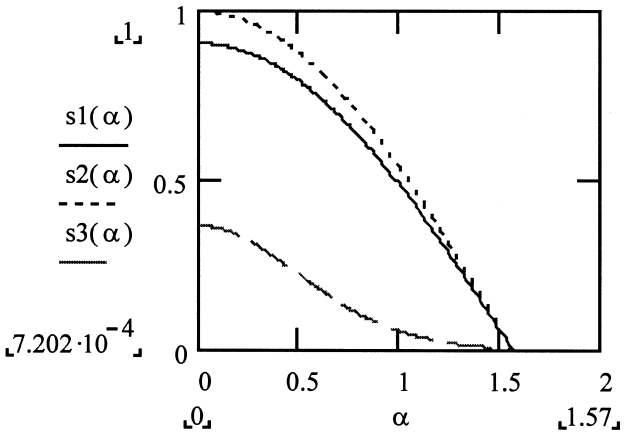

Fig. 5. Comparisons of the similarity measures (III).

3.2. Impact of a on the similarity measure

Eq. (4) is used to generate Fig. 6.

The X- and Y-axes are the parameter a and similarity, respectively, $0 < a \leqslant 1$, $d' = 10$, $d = 30$, and $\alpha = \pi/4$. Fig. 6 shows that the similarity value is very sensitive when the parameter a ranges from 0.7 to 0.97. It implies that the parameter a should be selected in that range.

3.3. Impact of α on the similarity measure

Eq. (4) is used to generate Fig. 7. The X- and Y-axes are the angle α and similarity, respectively, $0 < \alpha \leqslant \pi/2$, $d' = 10$, $d = 18$, and $a = 0.8$. Fig. 7 shows that the similarity value decreases when the variable α increases. Due to the symmetry, the angle α ranges from zero to $\pi/2$.

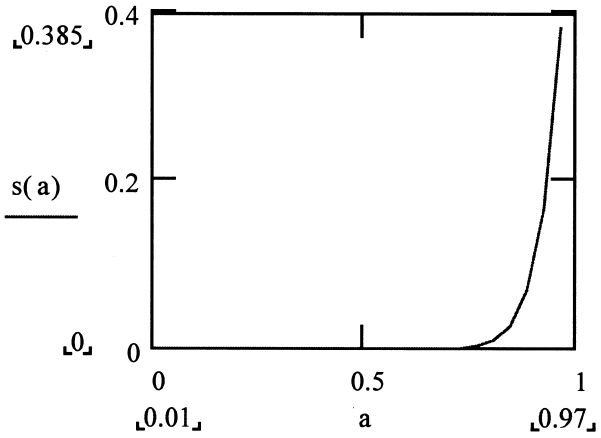

Fig. 6. Impact of a on the similarity measure.

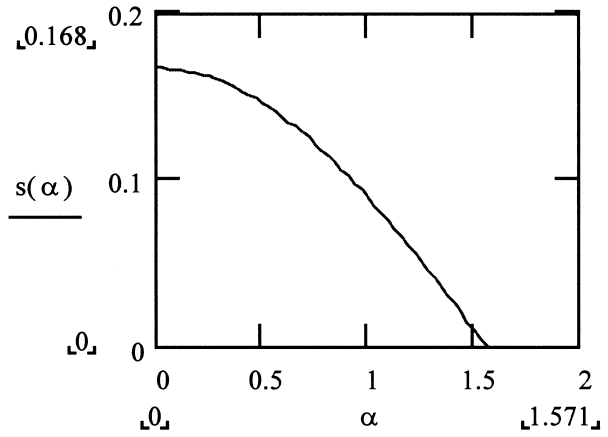

Fig. 7. Impact of α on the similarity measure.

3.4. Impact of d (d') on the similarity measure

Eq. (4) is used to generate Fig. 8. The X- and Y-axes are d and similarity, respectively, $d' = 10$, $\alpha = \pi/4$ and $a = 0.99$. There is a maximum similarity value in Fig. 8. Fig. 8 shows that the similarity value decreases when the variable d is larger than the distance from the compared object to the origin of the document vector space.

The analysis of d' is similar to that of d because they are symmetric in the newly developed similarity measure.

3.5. Iso-similarity analysis of a

Iso-similarity analysis demonstrates the relationships between two parameters when similarity has different values.

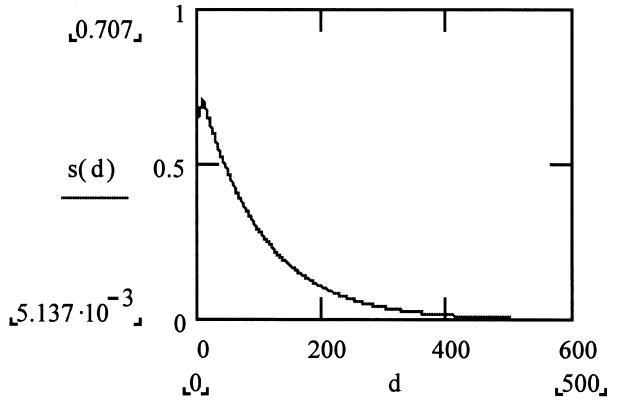

Fig. 8. Impact of d (d') on the similarity measure.

From Eq. (4):

$$a = \left(\frac{s}{\cos(\alpha)} \right)^{1/|d-d'|} \tag{16}$$

The iso-similarity analysis of a is generated by Eq. (16) (see Fig. 9). Where the X- and Y-axes are d and a, respectively, $a1(d)$, $a2(d)$, and $a3(d)$ correspond to $s = 0.1$, $s = 0.3$ and $s = 0.5$, respectively, $\alpha = \pi/4$, and $d' = 100$. The smaller the similarity value s, the smaller the corresponding parameter a.

3.6. Iso-similarity analysis of α

From Eq. (4):

$$\alpha = \arccos\left(\frac{s}{a^{|d-d'|}} \right) \tag{17}$$

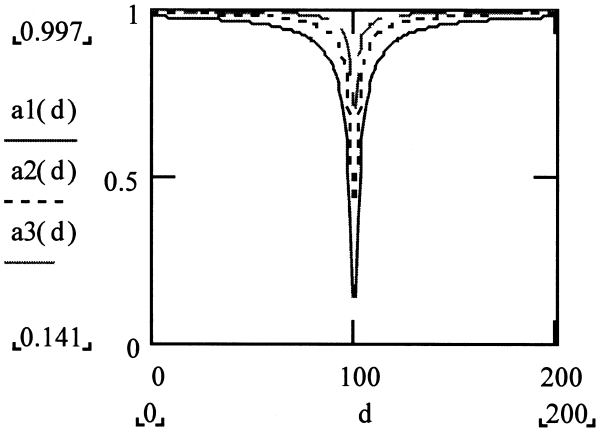

Fig. 9. Iso-similarity analysis of a.

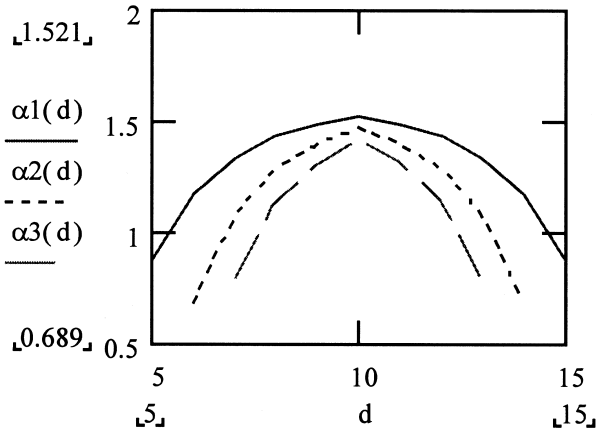

Fig. 10. Iso-similarity analysis of α.

The iso-similarity analysis of α is generated by Eq. (17) (see Fig. 10). Where the X- and Y-axes are d and α, respectively, $\alpha 1(d)$, $\alpha 2(d)$, and $\alpha 3(d)$ correspond to $s = 0.05$, $s = 0.1$ and $s = 0.15$, respectively, $a = 0.6$ and $d' = 10$. The smaller the similarity value s, the larger the corresponding parameter α.

4. Conclusion

After the traditional distance-based similarity measure and angle-based similarity measure are analyzed, an important concept, the iso-extent contour in a document vector space, is presented. The concept is used to construct the newly developed similarity measure. Based on both the distance strength and the direction strength a newly developed similarity measure is introduced. The difference in iso-extent contour between two objects, $|d-d'|$, is employed to describe the contribution of the distance parameter, and the angle between them vis-à-vis the origin of the document vector space is applied to describe the contribution of direction. In fact the distance modifier $a|d-d'|$ is used to modify the traditional angle-based similarity measure.

The impact of the distance strength on the newly developed similarity measure can be controlled by setting a parameter a. The larger the parameter a, the stronger the impact on the similarity measure, vice versa. The parameter a is very sensitive to the similarity measure when it is close to one. The application of the parameter a makes the newly developed similarity measure more flexible.

The comparisons among the distance-based, angle-based, and newly developed similarity measures facilitate an understanding of the features of the similarity measure. The comparisons are made from three unique perspectives: the distance, the direction, and both the distance and the direction.

The impact of the parameter a, the angle α, and the distance from one compared object to the origin of the document vector space d (d') on the similarity measure were investigated. Two iso-similarity contour analyses (a vs d and α vs d) were made.

It is widely recognized that the distance-based similarity measure and the angle-based similarity measure have different strengths in predicting the relevance of two objects in vector space. They are not compatible. That is, adopting one similarity measure for its strengths means giving up the differing strengths of the other. The main contribution of this paper is to find a niche between the two similarity measures and to offer an integrated approach to keep the strengths of both.

The proposed similarity measure has the potential for use in a vector-based information retrieval systems as an independent similarity measure to predict relevance, and for use to determine document positions in a browsing retrieval environment like VIBE (Olsen et al., 1993; Olsen & Korfhage, 1994). It may bring a new dimension to document cluster analysis approaches in which distances between documents are a dominant factor.

The new concept Iso-Extent Contour was introduced in this paper to answer questions such as why two documents with the same Cosine values but different locations in the vector space cannot be effectively identified in the Cosine approach, and what the difference of location means in terms of similarity, and how their difference can reasonably be explained and measured. These questions are well-known but little has been known about the answers.

We realize that an experimental study for a newly developed similarity measure is important. However, the theoretical soundness of a similarity measure must first be shown. It is the foundation of a similarity measure and the first step it is development. This paper focuses on addressing the theoretical soundness and property analysis of the proposed similarity measure. We plan to conduct a follow-up experimental investigation to evaluate the performance of several proposed similarity measures: the distance-based similarity measure, the angle-based similarity measure, the integrated distance-angle measure proposed here and the measure proposed earlier (Zhang, 1999) in a visual retrieval environment. The results will evaluate the proposed similarity measure from a practical perspective.

References

Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1998). Optimizing similarity using multi-query relevance feedback. *Journal of the American Society for Information Science*, *49*(8), 742–761.

Croft, W., & Harper, D. (1979). Using probabilistic models of information retrieval without relevance information. *Journal of Documentation*, *35*, 285–295.

Cronin, B. (1994). Tiered citation and measures of document similarity. *Journal of the American Society for Information Science*, *45*(7), 537–538.

Ellis, D., Turner, H. J., & Willett, P. (1993). Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management*, *3*(2), 128–149.

Falkowski, B. J. (1998). On certain generalizations of inner product similarity measures. *Journal of the American Society for Information Science*, *49*(9), 854–858.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information retrieval: data structures and algorithms*. Englewood Cliffs, NJ: Prentice Hall.

Fricke, M. (1997). Information using likeness measures. *Journal of the American Society for Information Science*, *48*(10), 882–892.

Griffiths, A., Luckhurst, H. C., & Willett, P. (1986). Using document similarity information in document retrieval systems. *Journal of the American Society for Information Science*, *37*(1), 3–11.

Hamers, L., Hemeryck, Y., & Herweyers, G. (1989). Similarity measures in scientometric research: the Jaccard index vs Salton's cosine formula. *Information Processing and Management*, *25*(3), 315–318.

Jones, W. P., & Furnes, G. W. (1987). Pictures of relevance: a geometric analysis of similarity measures. *Journal of the American Society for Information Science*, *38*(6), 420–442.

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

Kwok, K. L. (1985). A probabilistic theory of indexing and similarity measure based on cited and citing documents. *Journal of the American Society for Information Science*, *36*(5), 242–351.

McGill, M., Koll, M., & Noreault, T. (1979). An evaluation of factors affecting document ranking by information retrieval systems. School of Information Studies, Syracuse University, Syracuse, NY.

Meadow, C. T. (1992). *Text information retrieval systems*. San Diego, CA: Academic Press.

Nuchprayyoon, A. (1996). GUIDO: a usability study of its basic information retrieval operations. Ph.D. dissertation, School of Information Sciences, University of Pittsburgh.

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

Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., & Williams, J. G. (1993). Visualization of a document collection: the VIBE system. *Information Processing and Management*, *29*(1), 69–81.

Radecki, T. (1982). On a probabilistic approach to determining the similarity between Boolean search request formulations. *Journal of Documentation*, *38*(1), 14–28.

Radecki, T. (1985). A theoretical framework for defining similarity measures for Boolean search request formulations, including some experimental results. *Information Processing and Management*, *21*(6), 501–524.

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of searching terms. *Journal of the American Society for Information Science*, *27*, 129–146.

Robertson, S. E., & Walker, S. (1997). On relevance weights with little relevance information. *Proceedings of the Twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 16–24). Philadelphia, PA: ACM.

Rousseau, R. (1998). Jaccard similarity leads to the Marczewski–Steinhaus topology for information retrieval. *Information Processing and Management*, *34*(1), 87–94.

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. New York: Addison-Wesley.

Travison, D. (1987). Term co-occurence in cited/citing journal articles as a measure of document similarity. *Information Processing and Management*, *23*(3), 183–194.

Tudhope, D., & Taylor, C. (1996). A unified similarity coefficient for navigating through multi-dimensional information. *Proceedings of the 59th Annual Meeting of the American Society for Information Science* (pp. 67–70). Medford, NJ: Information Today, Inc.

Tudhope, D., & Taylor, C. (1997). Navigation via similarity: automatic linking based on semantic closeness. *Information Processing and Management*, *33*(2), 233–242.

Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.

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

Zhang, J., & Korfhage, R. R. (1999). A distance and angle similarity measure. *Journal of the American Society for Information Science*, *50*(9), 772–778.