A Distance and Angle Similarity Measure Method

Jin Zhang* and Robert R. Korfhage†
School of Information Sciences, University of Pittsburgh, 135 N. Bellefield Avenue, Pittsburgh, PA 15260.
E-mail: jzhang@sis.pitt.edu

This article presents a distance and angle similarity measure. The integrated similarity measure takes the strengths of both the distance and direction of measured documents into account. This article analyzes the features of the similarity measure by comparing it with the traditional distance-based similarity measure and the cosine measure, providing the iso-similarity contour, investigating the impacts of the parameters and variables on the new similarity measure. It also gives the further research issues on the topic.

Introduction

The similarity measure is an essential concept in information retrieval. It is widely used to judge whether a document matches a query, or to measure the similarity of two documents. In other words, the similarity measure allows a user to arrange or exhibit retrieved documents in decreasing order of similarity with respect to the query; to downsize a retrieved set by removing the documents with lesser similarity to the query; to measure discriminative value of indexing terms; and to dynamically adjust the retrieval strategy by adding more terms with high similarity and discarding the terms with low similarity. Furthermore, similarity measures can be applied to construct visualization interfaces to facilitate information retrieval.

A good similarity measure is an important factor that contributes to satisfactory precision and recall ratios in information retrieval.

Different information retrieval systems usually take different similarity measures. The distance and angle integrated similarity measure introduced here is a vector-based similarity measure.

According to McGill et al. (1979), there are more than 60 different similarity measures. These include the inner product, Dice coefficient, cosine coefficient, Jaccard coefficient, overlap coefficient (Frakes & Baeza-Yates, 1992; Korfhage, 1997; Meadow, 1992; Salton, 1968, 1989), the spreading activation similarity measure (Jones & Furnes, 1987), and some probability-based similarity measures (Croft & Harper, 1979; Kwok, 1985; Robertson & Sparck, 1976; Van Rijsbergen, 1979; as well as Robertson & Walker, 1997). Among them, the most popular are the distance-based similarity measure and the angle-based cosine measure.

Each similarity measure has its strengths and weaknesses in practice. Although much research has been done on similarity measures, the combination of different similarity measures is rarely considered. Research on the combination of different similarity measures has the potential to provide a new and unique approach to similarity research. The distance and angle integrated similarity measure attempts to organically combine a distance-based similarity measure with the angle-based cosine measure, to take advantage of the strengths of both and to make similarity measurement more scientific and accurate.

Fundamental, Features, and Analysis

To better understand the rationale of the proposed distance and angle integrated similarity measure, we should analyze the strengths and weaknesses of both the angle-based cosine measure and distance-based measure, from which the new idea shall be elicited.

The angle-based cosine measure is a direction-based similarity measure. It measures the similarity between a reference point and a document based only on the direction of the document in the document space vis-à-vis the reference point and the origin of the coordinate, ignoring the impact of the distance between the reference point and the document. The cosine measure can effectively identify documents in a vector document space that have the same indexing term distribution within the each document; that is, they have the same indexing terms, the same proportion of weights of any pair of indexing terms between two documents. This characteristic can be employed to identify documents with the same subject but at different levels in a document vector space.

* To whom all correspondence should be addressed. Current address: Jin Zhang, School of Library and Information Science, University of Wisconsin–Milwaukee, 2400 E. Hartford Ave., Milwaukee, WI 53211.

† Deceased.

Received July 17, 1998; revised March 8, 1999; accepted March 8, 1999.

Suppose R is a reference point, Di, Dj are two documents:

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

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

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

$x_{jr} = c^{*}x_{ir}$, $(r = 1, \ldots, n)$, c is a constant.

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

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

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

$$\text{cosine}(R, Dj) = \text{cosine}(R, Di) \quad (2)$$

Therefore, if Di and Dj are similar to R in terms of the cosine measure, and if the weights of the indexing terms are associated with the frequencies of the terms in the documents, the difference between the two documents will only be affected by measures of the lengths of the documents. The essence of the cosine measure is that it can identify the documents in terms of the indexing term distribution.

From the analysis we know that the direction of a document in a document vector space does affect its similarity to a certain object. However, it is not the only factor that can influence its similarity.

On the other hand, in the distance-based similarity measure, the similarity can be transformed from the distance between the document and the reference point as follows:

$$s = a^{-d} \quad (3)$$

where d is the distance, and a is a constant whose value is greater than 1.

The distance-based similarity measure follows the philosophy that documents close together are likely to be highly similar. In this case, all directions are considered equal.

The distance-based similarity measure takes only the impact of the distance into account, regardless of the direction of the document. In other words, documents with the same distance to the reference point shall have the same similarity. This approach can resolve the inherent weakness of the cosine measure that is that it cannot distinguish documents that have the same direction, but are far from each other in terms of distance in a document space. Although two documents share the same direction, it is argued that the validity of the high similarity of two documents is reduced to some extent when they are far apart in terms of distance. For example, there are three documents $D1$, $D2$, and $D3$ with the same direction in a document vector space (see Fig. 1).

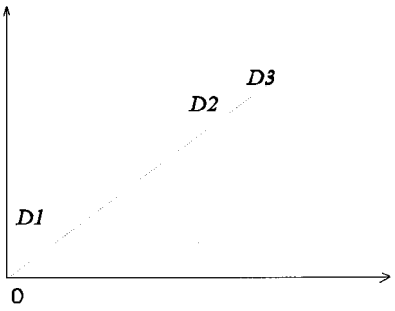

FIG. 1. Three documents with the same direction.

Because documents in the same direction vis-à-vis the origin of a document vector space have the same keyword distribution with proportional weights, the differences among these documents are reflected the extent to which they address the same topic. The similarity between $D1$ and $D3$ is the same as that between $D2$ and $D3$ in terms of the angle-based similarity measure. Similarity between two objects should be measured by both the topics they address and the extent to which they address. It is clear that document $D2$ should be more similar to $D3$ than $D1$ because document $D2$ addresses the same topic in more detail than document $D1$. Obviously, the farther apart documents $D1$ and $D2$ are, the bigger the difference should be. However, due to the inherent weakness of the angle-based similarity measure, ignoring the impact of the distance on the similarity, it cannot discern the difference in measuring the similarities of a group of documents with the same direction in a document vector space.

It is possible that two documents are quite similar in terms of the distance-based similarity measure but they are absolutely not similar in terms of the angle-based similarity measure.

Once a query, a distance-based similarity measure, and a document are selected, a contour is defined and documents within the contour are more relevant than that document. We can get a nice, symmetric "mountain" of relevant documents with the most relevant nearest the query. What modification by an angle-based similarity measure dose is to contour the surface of this mountain, depressing it more in some places than others, because it is quite likely that a document within the contour is less relevant than that document in terms of angle-based similarity measure.

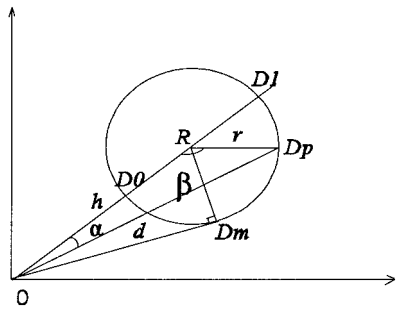

FIG. 2. Change of distance-based similarity measure in a document space.

The above analysis demonstrates that the two traditional similarity measures partially reflect the similarity of the compared objects, and they are complementary with respect to the distance and direction.

The complementary feature of two measures suggests that it would be useful to develop a new similarity measure, taking advantages of both and discarding the disadvantages. It is this aim that the new distance and angle integrated similarity measure attempts to achieve.

Once the distance between a reference point and a document is fixed, the similarity vis-à-vis the reference point is also fixed with respect to the distance-based similarity measure. In fact, the distance and the reference point will determine a hypersphere in the document space: the center of the hypersphere is the reference point, the radius is the distance between the reference point and the document. In this instance, each document in the circle has another similarity measure, vis-à-vis the axis, formed by the reference point and the origin in terms of the cosine measure. The similarity of the documents on the circle varies with different positions; in most the cases, they are not equal. When a document is located at the intersections between the axis and the circle (there are two such points—*D*0 and *D*1), the cosine measure of the document reaches the maximum value 1. The minimum value of the similarity depends on the length between the reference point and the document (see Fig. 2).

The phenomenon suggests that we could use the change of the direction-based similarity measure when a document moves along the circle to modify the distance-based similarity measure so that the new similarity could reflect not only the contribution of the distance, but also the contribution from the direction of the measured document. This is the rationale for the combined distance and angle similarity measure method.

In Figure 2, *r* is the radius of the circle, *R* is the reference point, *D*0, *D*1, *Dp*, and *Dm* are the documents in the circle, *h* is the distance between *R* and *O*, *d* is the distance between *Dp* and O. α and β are the angles formed by *RO* and *DpO*, *DpR*, respectively.

Once *Dp* is selected and the values of *r* and *h* are fixed, the maximum value of α is:

$$\alpha_{\max} = \arcsin(r/h) \tag{4}$$

If *Dp* is any point in the circle, the corresponding α is:

$$\alpha = \arccos \frac{\sum_{i=1}^{n} x_{1i} * x_{2i}}{\left(\sum_{i=1}^{n} x_{1i}^2\right)^{1/2}\left(\sum_{i=1}^{n} x_{2i}^2\right)^{1/2}} \tag{5}$$

where $Dp(x_{11}, x_{12}, \ldots, x_{1n})$; $R(x_{21}, x_{22}, \ldots, x_{2n})$; and *n* is the dimensionality of the document space.

The new distance and angle similarity measure is defined as follows:

$$s = a^{-r} * c^{k} \tag{6}$$

The effect of the parameters *a* and *c* will be discussed in detail later.

$$k = \frac{\alpha}{\alpha_{\max}}$$

$$k = \arccos \frac{\sum_{i=1}^{n} x_{1i} * x_{2i}}{\left(\sum_{i=1}^{n} x_{1i}^2\right)^{1/2}\left(\sum_{i=1}^{n} x_{2i}^2\right)^{1/2}} * \frac{1}{\arcsin(r/h)}$$

$$k = \arccos \frac{\sum_{i=1}^{n} x_{1i} * x_{2i}}{d * h} * \frac{1}{\arcsin(r/h)} \tag{7}$$

In fact, a^{-r} is the distance-based similarity measure, c^k is a modifier, where *k* is defined in Equation (7).

To maintain *s* in the 0 to 1 range, we require $0 < c \leq 1$. The effect of the parameters *a* and *c* will be discussed in detail in the section on The Effects of Parameters *a* and *c* on the Similarity Measures.

Note that in Equation (7) the angles rather than cosine values are used to express the impact of the direction rather than their cosine values, which reduces the complexity of the computation, and simultaneously keeps its basic characteristics. Because the document space vector elements are nonnegative, *R* is always in the first quadrant in the vector document space. We assume that *r* is sufficiently small that the circle lies entirely in the first quadrant.

Equations (6) and (7) show that the value of the new similarity measure shall be between 0 and 1.

Because the new similarity measure takes the angle and the distance into consideration, the problem discussed in Figure 1 can be avoided. If documents *D*1, *D*2, and *D*3 have the same direction in a document vector space, their corresponding angle α should be equal to zero in the new similarity measure. It suggests that the similarity measure is $s = a^{-r}$ when measured documents have the same angle with a reference point. It means that the similarity between the documents *D*1 and *D*3 is different from that between the

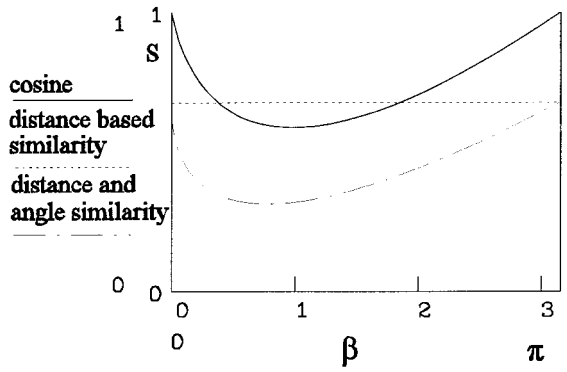

FIG. 3. Relationships among angle-based, distance-based, and distance-angle integrated similarity measures.

documents *D*2 and *D*3 due to the different *r* values. The difference between the two similarities depends on the distance between *D*1 and *D*3 as well as the distance between *D*2 and *D*3. In other words, the extent to which they address the same topic is reflected in the new similarity measure.

Now consider the features of the new similarity measure in detail.

To illustrate the impact, the entire circle should be displayed. However, for simplicity of the display and due to the symmetry of the circle vis-à-vis *RO*, only one-half of the circle is displayed, that is, β ranges from zero to π.

The Relationship of the New Similarity Measure to the Distance-Based Similarity Measure and the Cosine Measure

In Figure 2, when any document *Dp* moves from *D*0 to *D*1 along the circle the similarity varies with different similarity measures. For the distance-based similarity measure, it is a constant, depending on the distance between *Dp* and *R*, (See Fig. 3); for the cosine measure, as *Dp* moves from *D*0 to *D*1, it decreases from D0 to *Dm*, then increases from *Dm* to *D*1. The minimum value is cos[arcsin(r/h)], the maximum value is 1 (see Fig. 3); the new similarity measure has characteristics of both similarity measures: first, it is changeable; second, it has a minimum value at the same position as the cosine measure does. This value is smaller than that of the cosine measure; finally, its maximum value is equal to that of the distance-based similarity measure (see Fig. 3). In Figure 3, *x*-axis is the angle and *y*-axis is the similarity.

The Impact of the Parameter r *on the Similarity Measure*

Suppose that *r* changes, but the center of the circle is stationary; in other words, *h* is fixed.

Equation (13) is used to generate Figure 4.

In Figure 4, $h = 9$, $a = 1.11$, and $c = 0.5$.

The four curves are associated with $r = 1$, 3, 5, and 7, respectively, in Figure 4.

Notice that when *r* changes, the maximum and minimum values of the similarity measure, and the position of the minimum value in the *X*-axis also change. The smaller the value of *r*, i.e., the nearer the document to the reference point, the higher the similarity value, and vice versa.

The Impact of the Parameter c *on the Similarity Measure*

From Figure 2, we have:

$$r * \sin\beta = d * \sin\alpha \tag{8}$$

$$r * \cos\beta + d * \cos\alpha = h \tag{9}$$

From Equations (8) and (9):

$$r * \cos\beta + r * \frac{\sin\beta}{\sin\alpha} * \cos\alpha = h$$

$$\alpha = \arctan\left(\frac{\sin\beta}{h/r - \cos\beta}\right) \tag{10}$$

From Equations (6), (7), and (10):

$$s = a^{-r} * c^{\{\arctan[\sin\beta/((h/r)-\cos\beta)]/\alpha_{\max}\}} \tag{11}$$

The impact of the parameter *c* on the similarity measure is illustrated in Figure 5, where $h = 5$, $a = 1.11$, $r = 3$, $\alpha_{\max} = \pi/6$. The four curves are associated with $c = 0.2$, 0.4, 0.6, and 0.8, respectively. Notice that when $c = 1$, Equation (6) becomes the distance-based similarity measure. Figure 5 indicates that the smaller the value of *c*, the greater the influence of c^k as a modifier on the similarity measure, and vice versa. Each curve reaches its minimum value at the same position. They have the same maximum value a^{-r}.

The Impact of the Parameter a *on the Similarity Measure*

Equation (11) yields Figure 6, where values of *h*, *r*, and $\alpha_{\max}$ are the same as above, $c = 0.5$.

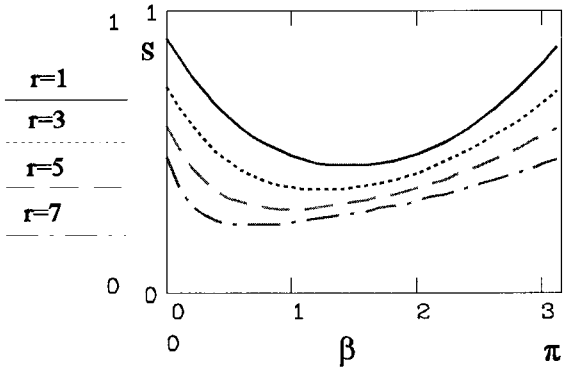

FIG. 4. Impact of *r* on the similarity measure.

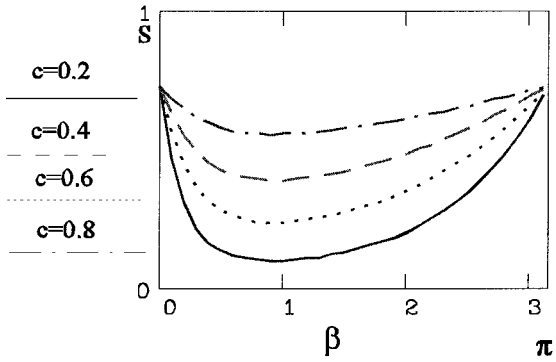

FIG. 5. Impact of c on the similarity measure.

The four curves are associated with $a = 1.1$, 1.3, 1.5, and 1.7 respectively. The smaller the value of a, the greater the similarity, and the greater the difference between the minimum and the maximum. Each a determines different minimum and maximum similarity values, but the position at which the curves reach their minimum points is same.

The Impact of the Parameter α_{max} on the Similarity Measure

Suppose as $\alpha_{\max}$ changes, the center of the circle does not move, i.e., h stays the same. The change of $\alpha_{\max}$ will then affect the radius r; r is a function of the $\alpha_{\max}$; thus:

$$s = a^{(-h*\sin\alpha_{\max})} * c^{\{\arctan[\sin\beta/((1/\sin\alpha_{\max})-\cos\beta)]/\alpha_{\max}\}} \quad (12)$$

The four curves ($\alpha_{\max} = \pi 3$, $\pi/4$, $\pi/5$, and $\pi/6$) are presented in Figure 7. The parameter values are $h = 5$, $a = 1.11$, and $c = 0.5$.

When $\alpha_{\max}$ changes , the positions at which the curves achieve minimum values vary with the different $\alpha_{\max}$. The smaller the $\alpha_{\max}$, the smaller the minimum value, and vice versa.

The Impact of the Parameter h *on the Similarity Measure*

When the value of h changes, the radius of the circle does not change. The change would affect $\alpha_{\max}$ [see Equation (4)]; therefore:

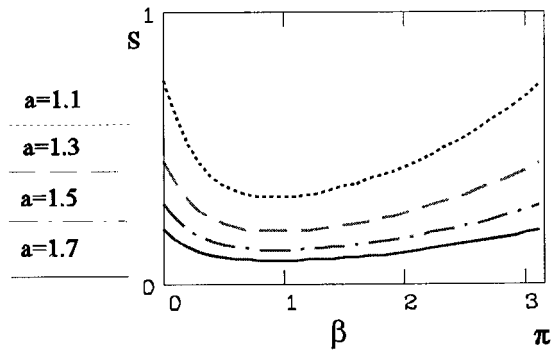

FIG. 6. Impact of a on the similarity measure.

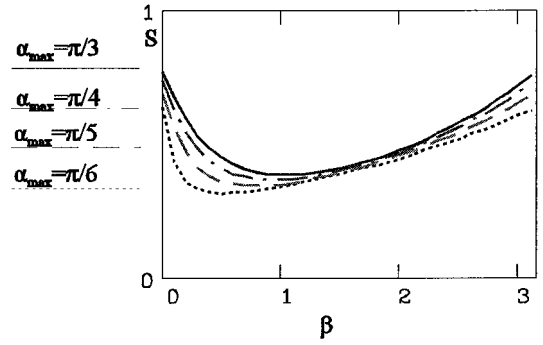

FIG. 7. Impact of $\alpha_{\max}$ on the similarity measure.

$$s = a^{-r} * c^{\{\arctan[\sin\beta/((h/r)-\cos\beta)]/\arcsin(r/h)\}} \quad (13)$$

The four curves with $h = 3$, 5, 7, and 9, respectively, are shown in Figure 8, where $a = 1.11$, $r = 2$, $c = 0.5$. As value of h changes, the positions at which the curves gain their minimum values also change, the corresponding values change, but the maximum value does not. The smaller the value of h, the larger the minimum value; and vice versa.

Iso-similarity Contour Analysis

Iso-similarity contour with respect to the parameter c

From Equation (11):

$$\log s = \log(a^{-r}) + \arctan\left(\frac{\sin\beta}{h/r - \cos\beta}\right) * \frac{\log c}{\alpha_{\max}}$$

$$c = 10^{[\alpha_{\max} * \log(s * a^r)]/\{\arctan[\sin\beta/((h/r)-\cos\beta)]\}} \quad (14)$$

The four contours ($s = 0.1$, 0.3, 0.5, and 0.7) are given in Figure 9, where $h = 5$, $a = 1.11$, $\alpha_{\max} = \pi/6$.

Figure 9 shows that the smaller the value of s, the lower the iso-similarity contour; and vice versa.

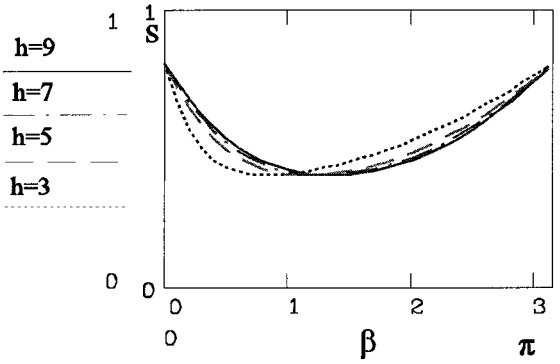

FIG. 8. Impact of h on the similarity measure.

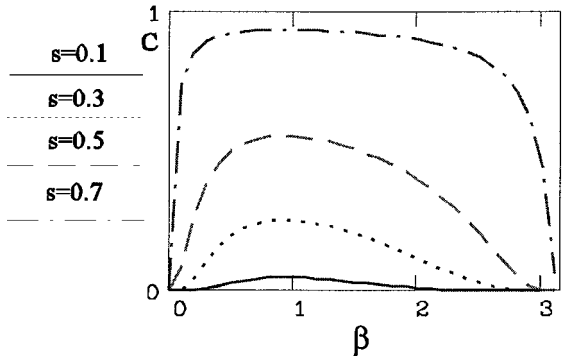

FIG. 9. Iso-similarity analysis of *c*.

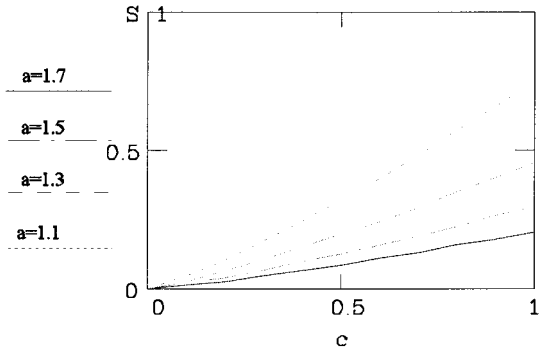

FIG. 11. Effect of c on the similarity measure.

Iso-similarity contour with respect to the parameter a

From Equation (11):

$$a^r = \frac{1}{s} * c^{\{\arctan[\sin\beta/((h/r)-\cos\beta)]/\alpha_{\max}\}}$$

$$a = 10^{(1/r)*\log(s^{-1}*c^{\arctan[\sin\beta/((h/r)-\cos\beta)]})} \quad (15)$$

The four contours with $s = 0.1$, 0.3, 0.5, and 0.7 are exhibited in Figure 10, where $h = 5$, $\alpha_{\max} = \pi/6$, $c = 0.8$.

Figure 10 indicates that the smaller the value of *s*, the higher the contour.

The Effects of Parameters a *and* c *on the Similarity Measures*

The parameters *a* and *c* are artificial, affecting the display of similarity. We note that *a* is related to the distance *r*, and *c* is related to the angle β. The two display parameters are restricted in range: $a > 1$, and $0 < c \leq 1$. When either *a* or *c* is set to 1, the similarity measure is independent of the associated document parameter. To show the interaction between *a* or *c*, we assume a hypothetical query and document, thus fixing the parameters *h*, *r*, and β.

Let us discuss the effect of parameter *a* on the similarity measure when *c* changes.

According to Equation (11), the four contours with $a = 1.1$, 1.3, 1.5, and 1.7 are generated in Figure 11, where $h = 5$, $\alpha_{\max} = \pi/6$, $r = 3$, $\beta = \pi/3$, and *c* from 0 to 1.

Figure 11 shows that for a fixed *a*, when *c* increases, the corresponding similarity value increases. The lower the value of *a* is, the more the similarity value increases.

The effect of parameter *c* on the similarity measure when *a* changes is as follows.

According to Equation (11), the four contours with $c = 0.2$, 0.4, 0.6, and 0.8 are given in Figure 12,where $h = 5$, $\alpha_{\max} = \pi/6$, $r = 3$, $\beta = \pi/3$, and *a* from 1 to 3.

Figure 12 demonstrates that when *a* increases, the similarity value of *c* decreases quickly to zero.

Figures 11 and 12 show that there is a range of *a* and *c* values that will yield this similarity value. The user's choice for these display parameters will reflect his emphasis on distance (low *a* value) or angle (high *c* value) as the dominant similarity factor.

Conclusion

The distance and angle similarity measure presents a new approach to integrating both a distance-based similarity measure and a direction-based similarity measure. It takes the effects of the distance and direction of documents on the similarity measure into account. The contributions of both the distance and angle to the similarity value are adjustable by controlling the corresponding parameters.

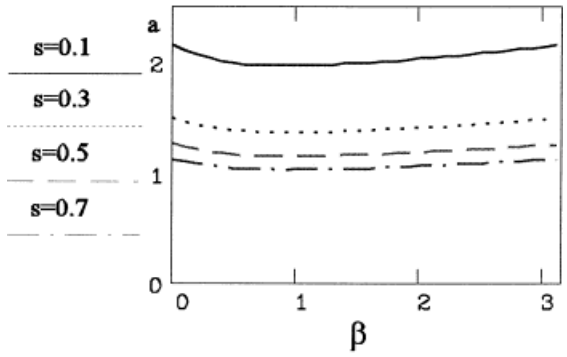

FIG. 10. Iso-similarity anaysis of *a*.

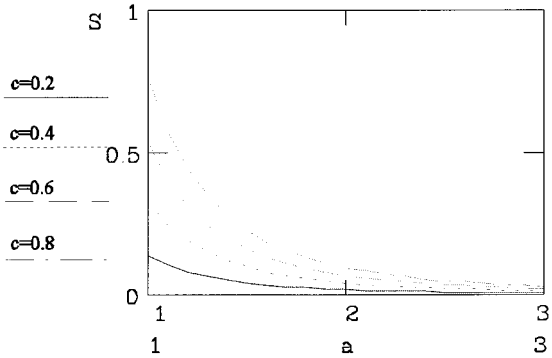

FIG. 12. Effect of *a* on the similarity measure.

The analysis of the parameters such as a, c, α_{max}, r, and h in the similarity measure tells the users how to apply the similarity measure appropriately: the parameter c could be used to control the strength of the direction of the measured documents. The effect of the direction on this similarity measure is based on the strength of the distance of the document. The parameter a is applied to adjust the strength of the distance. The parameter h, which is the distance from the origin to the reference point, is indirectly associated with the impact of the direction. The parameter r, which is the distance from a document to the reference point, influences the strength of both distance and direction; it is one of the key variables in the similarity measure, as it also determines the maximum and minimum values of the similarity measure values.

The iso-similarity analysis shows that the value of c, and to a lesser extent the value of a, impact the perceived similarity value of a document. This could help users to select the parameters to best advantage.

The analysis of effects of interaction between two parameters a and c on the similarity measures presents more useful information for the selection of a and c.

The way of measuring the angle of a document influences the determination of the maximum angle α_{max}. It is important when this new similarity measure is applied to the distance–angle-based visual environment.

Basically, in this similarity measure the four parameters can be classified into two groups. Group 1 contains two parameters relating a document and a query (reference point) positions ("h" and "r") in a document vector space, they are determined only by the document and the query, not by any similarity measure, and they impact a similarity measure. Group 2 contains parameters relating this new similarity measure ("a" and "c"); users are allowed to manipulate them to control the impact of distance and angle on the similarity measure, and they are determined by users rather than the document and the query.

In the new distance and angle integrated similarity measure the distance-based measure is taken as the primary one, and it is reduced by the angle-based similarity measure when the maximum similarity value is used as the compared object (starting point is $D0$ or $D1$ in Fig. 2). However, in the same situation when the minium similarity value is used as the compared object (different starting point Dm in a vector space in Fig. 2), it is increased rather than decreased by the angle-based similarity measure.

Directions for further research include integrating other distance-based similarity measures with the direction-based similarity measure, for instance, substituting $a^{-r\times r}$ for a^{-r} in the distance and angle similarity measure; coordinating the use of the different parameters, etc. This article only focuses on discussing the properties of the new similarity measure. It is necessary to conduct an experimental study to investigate the performance among the distance-based similarity measure, the angle-based similarity measure, and this new similarity measure in future research, allowing people to understand the new similarity measure from a different perspective.

References

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

Frakes, W.B., & Baeza-Yates, R., Eds. (1992). Information retrieval: Data structure and algorithms, Englewood Cliffs, NJ: Prentice Hall.

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 Computer Pub.

Kwok, K.L. (1985). A probabilistic theory of indexing and ximilarity 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, Syracuse, NY: School of Information Studies, Syracuse University.

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

Robertson, S.E., & Sparck, J.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. In Proceedings of the Twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 16–24), Philadelphia, PA: ACM.

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.

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