An experimental investigation on the distance angle integrated similarity measure

Une étude expérimentale de la mesure de similarité intégrée basée sur l'angle de distance

Jin Zhang
School of Information Studies
University of Wisconsin–Milwaukee
Milwaukee, WI 53211
jzhang@uwm.edu

Résumé : Une étude expérimentale a été menée afin d'examiner et de tester les performances de repérage de la mesure de similarité intégrée basée sur l'angle de distance présentée dans un article antérieur. La performance de repérage de la nouvelle mesure de similarité basée sur l'angle, la distance, la conjonction, la disjonction, et les modèles elliptiques de repérage a été respectivement comparée à la performance de repérage de la mesure de similarité traditionnelle et à celle de la mesure de similarité basée sur l'angle. Les résultats démontrent que la mesure intégrée de similarité basée sur l'angle de distance offre une performance satisfaisante.

Abstract: An experimental study was conducted to examine and test retrieval performance of the distance angle integrated similarity measure introduced in an earlier paper. Retrieval performance of the new similarity measure within the angle, distance, conjunction, disjunction, and ellipse retrieval models was compared against retrieval performance of the traditional distance similarity measure and the angle similarity measure, respectively. The results demonstrate the distance angle integrated similarity measure achieved satisfactory performance.

Introduction

Similarity measure is essential and fundamental for information retrieval. Research on similarity measure has been a longstanding topic. Similarity measure is used to identify relevant objects such as a query and a document in an information retrieval system, to distinguish document clusters in a document collection, or to apply to automatic classification, categorization,

clustering (Na, Kang, and Lee 2007), automatic feedback (Chen and Fu 2007), and other applications such as visualization for information retrieval, such as VIBE (Olsen et al. 1993; Olsen and Korfhage 1994), GUIDO (Nuchprayyoon 1996), DARE (Zhang and Korfhage 1999a; Zhang 2000), TOFIR (Zhang 2000), and Lighthouse (Allan et al. 2001).

It is clear that many factors can affect similarity between two objects. Basic characteristics of a similarity measure depend primarily on which factors are identified and considered in its implementation. In other words, similarity measures can be developed from different perspectives or emphases, such as from the perspective of probability (Croft and Harper 1979; Radecki 1982; Kwok 1985; Salton 1989; Robertson and Walker 1997; Torvik, Weeber, and Swanson 2005), citation (Kwok 1985; Travison 1987; Cronin 1994), hyperlink (Calado, Cristo, and Goncalves 2006), fuzzy theory (Egghe and Michel 2003), search (Robertson and Sparck Jones 1976; Radecki 1985; Bartell, Cottrell, and Belew 1998; Kim and Choi 1999), and other meaningful perspectives (Ellis, Turner, and Willett 1993; Fricke 1997; Tudhope and Taylor 1997; Falkowski 1998; Rousseau 1998; Atlam et al. 2000; Atlam, Fuketa, and Morita 2003; Watters and Wang 2000; Michel 2001; Burrell 2005; Egghe 2006). Each similarity measure has its strengths and weaknesses. Different similarity measures may work for different tasks in different situations. It is the diversity of similarity measures that provides a wide selection spectrum for their application.

Because there are a variety of options for application of similarity measure, research on comparison and evaluation of similarity measures (McGill, Koll, and Noreault 1979; Kwok 1985; Griffiths, Luckhurst, and Willett 1986; Hamers, Hemeryck, and Herweyers 1989; Ellis, Turner, and Willett 1993; Qin 2000; Zhang and Rasmussen 2001; 2002) become crucial for users to understand the nature and feature of similarity measures, and that is important for them in order to make a correct decision on similarity measure selections.

It is widely recognized that both distance and direction characteristics of an object in a high dimensional document space can contribute to its similarity while the traditional distance-based similarity measure takes only the distance characteristic into consideration (Zhang and Rasmussen 2001). As a result, documents with the same distance-based similarity value may correspond to different directions. These differences are not obviously reflected in the traditional distance similarity measure. In an earlier paper (Zhang and Korfhage 1999b), a distance angle integrated

similarity measure was introduced. This measure utilizes the direction strength of a measured object in a vector-based information space based on a regular distance similarity measure. In this case, documents with the same distance-based similarity value but different directions are treated differently when their similarities are measured. In addition to the distance from a measured document to a query, the newly developed similarity measure depends on the distance between the query and the origin of the document space, and the distance between the document and the origin. The degree to which either the direction or the distance of a measured document affects the final similarity measure can be controlled and manipulated by users.

A detailed theoretical proof and property analysis of this similarity measure has been addressed (Zhang and Korfhage 1999b). The theoretical analyses imply that the distance angle integrated similarity measure should have more advantages than both the traditional angle similarity measure and distance similarity measure because both the angle strength and the distance strength are taken into consideration in the new similarity measure. However, theoretical soundness should be confirmed by experimental evidence. Theoretical expectations may differ from practical findings for some unpredictable reasons. Some unexpected variables may play a crucial role in the similarity measure. Consideration of direction in a distance-based similarity measure may have little or even no impact on the similarity measure, or experimental findings may show the opposite to the expectations. In other words, whether the new measure can achieve better retrieval performance than both the distance and angle similarity measures in practice is still a mystery. In the previous paper (Zhang and Korfhage 1999b), the authors pointed out that an experimental study should be conducted to examine the soundness of the presented similarity measure from a practical perspective. This is the motivation of this research.

The primary aim of this experimental study, as a follow-up to the theoretical discussion of the distance angle integrated similarity measure, is to investigate whether it achieves better retrieval performance than both the angle similarity measure and the distance similarity measure in practice.

Statement of the hypotheses

In order to better understand the paper, it is necessary to define some related concepts that will be used later.

Similarity measures

There are three different similarity measures involved in this experimental study: the distance similarity measure, the angle similarity measure, and the distance angle integrated similarity measure. Their definitions follow.

Suppose $Rk\ (x_{k1}, x_{k2}, \ldots, x_{kn})$ is a reference point (or query) (Korfhage 1997), and $Di\ (x_{i1}, x_{i2}, \ldots, x_{in})$ is a document in a document vector space. Here n is the dimensionality of the document space.

*SM*1(equation 1) is the distance similarity measure, where d is defined as the distance from a reference point Rk (query) to the measured Di (document), and g is a constant whose value is always greater than 1. If Rk and Di are the same, *SM1* is defined as 1. In a pilot study, the impact of the parameter g on the similarity was tested, and the testing data showed that when g got to 1.11, the similarity became more stable. So according to the pilot study, in this study g is set to 1.11.

$$SM1 = g^{-d}$$

$$SM1 = g^{-\left(\sum_{r=1}^{n}(X_{kr}-X_{ir})^2\right)^{1/2}} \tag{1}$$

*SM*2 (equation 2) is the angle similarity measure or the cosine measure, where α is an angle formed by the reference point Rk and the measured Di against the origin of the document space.

$$SM2 = \cos(\alpha)$$

$$SM2 = \frac{\sum_{r=1}^{n}(X_{kr} \times X_{ir})}{\left(\sum_{r=1}^{n} X_{kr}^{\,2}\right)^{1/2} \times \left(\sum_{r=1}^{n} X_{kr}^{\,2}\right)^{1/2}} \tag{2}$$

*SM*3 (equation 3) is the distance angle integrated similarity measure. The variables d, $d1$, and $d2$ are defined as the distance from the query Rk to the document Di (Rk and Di are not the same), the distance from the origin of the document s pace to Di, and the distance from the origin to Rk, respectively. The constant a is a constant whose value is greater than 1,

and constant c is larger than 0 but less than or equal to 1. The angle α_{max} is the maximum value that the variable α can reach. The detailed discussion on this similarity measure consults the previous study (Zhang and Korfhage 1999b).

$$SM3 = a^{-d} \times c^{\frac{\alpha}{\alpha_{max}}}$$

$$SM3 = a^{-d} \times c^{\arccos \frac{\sum_{p=1}^{n} x_{kp} \times x_{ip}}{d1 \times d2} * \frac{1}{\arcsin(d/d2)}} \quad (3)$$

Notice that the angle α is always smaller than or equal to angle α_{max}.

The previous study examined the impact of the parameters a and c on the similarity measure (Zhang and Korfhage 1999b). The similarity value is very active and sensitive when the parameter a ranges from 1.08 to 1.2 and the parameter c from 0.4 to 1. It implies that selection of these parameters in this experimental study should follow the principle. In order to achieve a satisfactory experimental result for the distance angle integrated similarity measure, the parameters a and c were set to 1/0.9 and 0.5 respectively.

The parameter c is used to control the impact of the direction/angle on the similarity measure. The valid range of the parameter c is from 0 to 1. The larger the parameter c is, the more impact the direction has on the similarity, and vice versa. In this study, the parameter c is set to 0.5. It implies that the direction would have an average impact on the similarity.

Information retrieval models

An information retrieval model (Korfhage 1997) offers a control mechanism for users to focus on a local area of interest in the document space. A retrieval model usually defines a contour in a document space. Documents within the contour are regarded as retrieved documents for that model. The shape of an information retrieval model contour varies and reflects its retrieval nature and emphasis. Its size and position in the document space are usually controllable.

Five information retrieval models were involved in this experimental study.

1. The angle (cosine) retrieval model determines a hyper cone in the high dimensional document space. The origin of the document space is always the point of the cone, and a specified angle (threshold) of the cone is used to control a retrieved area or documents. Increase (or decrease) of the angle (threshold) may result in increase (or decrease) in the number of retrieved documents.
2. The distance model corresponds to a hyper sphere in the high dimensional document space. Documents within the sphere are retrieved. The radius of the sphere, which can be manipulated by users, is the threshold for search. The centre of the sphere is the query (reference point).
3. The conjunction retrieval model defines an overlapping area of two contours in the document space. This model requires two reference points. Each reference point is associated with a contour like that of the distance retrieval model.
4. Similar to the conjunction retrieval model, the disjunction retrieval model also requires two reference points. However, the difference is that its contour consists of the two entire contours produced by the two reference points.
5. The ellipse retrieval model is also based on the two-reference-point-based retrieval model. In this model the sum of distance from a point on the contour to one reference point and distance from it to the other reference point maintains a constant no matter where a point is located on the contour. According to this definition, the contour of the ellipse retrieval model is symmetrical against the axis formed by the two reference points in the document space.

In summary, the four distance-based retrieval models (the distance model, conjunction model, disjunction model, and ellipse model) and one angle-based retrieval model were used in this study. Two of these five models are based on one reference point (the distance model and the angle model) and three are based on two reference points (the conjunction model, the disjunction model, and the ellipse model).

Proposed hypotheses

In order to examine the newly developed similarity measure from multiple perspectives, we tried to include the five information retrieval models,

ranging from retrieval models based on one reference point to models based on two reference points, and from an angle-based retrieval model to distance-based retrieval models. Within each individual information retrieval model, performance of the distance angle integrated similarity measure against that of both the angle similarity measure and distance similarity measure was compared respectively. In this case, performance of the new similarity measure in the contexts of different information retrieval models was examined and tested.

The newly developed similarity measure was compared with the distance similarity measure and angle similarity measure respectively in the experimental study, because the new measure evolves from the distance similarity measure and integrates the angle strength. Therefore, it possesses the distance and angle characteristics. The comparisons among the three similarity measures would help us to identify and understand which factor (angle, distance, or both) makes primary contribution to improvement of retrieval effectiveness.

The proposed hypotheses are stated as follows:

H1a There is no difference in retrieval performance among the three different similarity measures.

H1b There is no difference in retrieval performance among the five different information retrieval models.

H1c There is no interaction effect between similarity measure and information retrieval model.

H2 The distance angle integrated similarity measure within the angle retrieval model achieves better performance than the distance angle integrated similarity measure within all the distance-based retrieval models.

H3 The distance angle integrated similarity measure within the models based on one reference point achieves better performance than the distance angle integrated similarity measure within the models based on two reference points.

H4a The distance angle integrated similarity measure achieves better performance than the angle similarity measure within the angle retrieval model.

H4b The distance angle integrated similarity measure achieves better performance than the distance similarity measure within the angle retrieval model.

H5a The distance angle integrated similarity measure achieves better performance than the angle similarity measure within the distance-based retrieval models.

H5b The distance angle integrated similarity measure achieves better performance than the distance similarity measure within the distance-based retrieval models.

H1 assesses performance differences among the three similarity measures and their interaction among all information retrieval models. It gives users a comprehensive overview of their performance in this experimental study.

H2 and H3 examine performance of the new measure in two categories: angle-based retrieval model versus distance-based retrieval models, and models based on one reference point versus models based on two reference points. Note that the comparisons in these two hypotheses do not use the new similarity measure against either the distance-based similarity measure or the angle-based similarity measure. The comparisons are made between performance of the new similarity measures but in the different contexts of information retrieval models.

H4a and H4b (H5a and H5b) emphasize the comparisons between the new measure and the two traditional similarity measures within the angle-based retrieval model (the distance-based retrieval models). The comparisons are made between performances of two similarity measures in the same retrieval model context.

Methodology

The methodology employed in this experimental study is the same as that in a previous similar study (Zhang and Rasmussen 2002) which investigated the iso-content-based angle similarity measure within the same retrieval environment. Both the iso-content-based angle similarity measure and this similarity measure have something in common. Both utilize the strengths of distance and direction of a measured object to measure its similarity but in quite different ways.

The information retrieval system used in this experimental study is DARE (Distance Angle Retrieval Environment) application for Windows (Zhang and Korfhage 1999a; Zhang 2000). DARE is a visual information retrieval tool. It is equipped with the five information retrieval models and three similarity ranking options for retrieved documents previously discussed, as well as a visual browsing environment.

The database used in the experimental study is an Associated Press (AP) database containing full text of about 450 news reports from 1989. It came from the TREC database. Drug issues, global and local economy, the Eastern European political changes, area conflicts, entertainment news, and so on were primary topics during that time. Structured fields within each record include document number, headline, sub-headline, author, date, and full text.

The Minitab statistical analysis package was used for data analysis.

The participants were 32 graduate students in library and information science (see Appendix 2). All of them had experience searching an OPAC and the Internet and they understood basic concepts of information retrieval. They regularly used computers.

In this experimental study, recall was used to evaluate retrieval performance for each similarity measure.

Because in this study the retrieval results for the three similarity measures were determined in the post-processing step, the retrieval result sizes of the three similarity measures for a search were the same. It suggests that the precision comparison results among the three similarity measures would be similar to the recall comparison results among the three similarity measures. For the purpose of simplicity, only recall comparison was included in the study.

Although the recall values are not listed, because space is limited, all means and standard deviations of the recall values under different situations are listed from table 1 to table 11.

Query preparation. Ten queries for the five information retrieval models were prepared (see Appendix 1). There are two types of queries: those based on one reference point and those based on two. A statement for each query was attached so that subjects could better understand

the information need behind that query. Queries based on one reference point are used for the angle retrieval model and the distance retrieval model, while queries based on two reference points are used for the conjunction retrieval model, disjunction retrieval model, and the ellipse retrieval model.

Task preparation. Each individual search task included a query, query statement, assigned information retrieval model, instruction, and other requirements. Queries based on one reference point and two reference points were randomly assigned to the one-reference-point-based and two-reference-point-based information retrieval models, respectively.

Task assignment. Tasks were randomly distributed among subjects. Each subject was required to carry out at least one of the two query types. Eighty tasks were prepared and scattered among 32 participants. Two or three tasks were assigned to each individual subject so that 80 tasks could be evenly assigned to 32 subjects. Each information retrieval model was used 16 times in the study.

Search procedure. Each subject completed tasks assigned to him or her independently. For each task, a participant employed an assigned information retrieval model to search for documents in the database. An initial retrieved document set was produced, based on the retrieval model and requirement to include relevant documents for a query. Then the subject identified relevant documents from the initial retrieved document set, based on the specified information need statement. Finally, the relevant documents for a query were recoded and saved in a result file for later data analysis.

In this phase, subjects did not use any similarity measure to determine relevant documents from retrieved document sets. In other words, the search results judged by a subject were not ranked by any of the three similarity measures. The results were independent of the three similarity measures. The reason for not using any similarity measures was to avoid unnecessary experimental bias. If subjects had used the three similarity measures to rank the same predetermined retrieved document set and to pick up relevant documents from the same predetermined retrieved document set, they would have used the same predetermined retrieved document set multiple times. As a result, they would have become familiar with contents of documents in the predetermined retrieved document set, and that would have caused experimental bias. In that case, the second (third) round of selection of relevant documents for other similarity

measures may rely partially on their familiarity rather than their relevance. Another benefit of this strategy was that it also shortened the experimental time for each task. The simple and effective experimental plan would be helpful for subjects to easily follow experimental procedures, to concentrate on a key part of the experiment, and therefore to reduce possible operational errors. It is even more important that the collected data in the experimental study can be reused to test other similarity measures, just because the data collection process is not associated with any similarity measures at this point.

Post-processing. The relevant documents of a query for each individual similarity measure were determined via the following steps:

1. On the basis of the search log, used the same reference point(s), the same information retrieval model, and the same threshold(s) to search the same database, determining the same size predetermined retrieved document set.
2. Ranked all documents in a predetermined retrieved document set by a corresponding similarity measure: the distance similarity measure, the angle similarity measure, or the distance angel integrated similarity measure.
3. After a retrieved document set was ranked by a specified similarity measure, the most relevant documents, which were located in the top of the ranked list, were kept to form a most relevant document set for that specific similarity measure. The number of the kept documents was the same as the number of documents in the standard answer set for that query. This newly identified set was compared with a standard answer set for the query to find the overlapping parts between the two sets. The overlapping documents were regarded as final relevant documents for that similarity measure against that query.

Following the same procedure, a relevant document set for each of the three similarity measures was produced, and corresponding results were recorded.

After this processing, each final document set produced by a subject derived three new subsets: one for the distance similarity measure, one for the angle similarity measure, and one for the distance angle integrated similarity measure, respectively.

It is worth pointing out that the data processing for each experimental task, described in the section "Post-processing," was completed by the experimenter rather than subjects. It was conducted after the experiment was done.

In summary, in this study 10 queries were prepared, 32 subjects participated, 80 tasks were generated, the average number of tasks per subject was 2.5, each information retrieval model was used 16 times, and total number of tasks for the three similarity measures was 240.

Data analysis

After the experimental preparation, design, search, data post-processing, and data collection, the collected data were used to test the proposed hypotheses.

The significance level α in this experimental study for all hypotheses was set to .05.

There are two ways to judge a test result: the *p-value* approach and the test statistic approach. The former approach bases acceptance of a hypothesis on the condition p value > α (α = .05) at the 100 α % significance level if the comparison between two factors are set to be "equal," or the condition p value = < α at the 100 α % significance level if the comparison between two factors are set to be "greater than." The latter approach bases the decision on whether a hypothesis is rejected or accepted on both a critical value and a significance level α.

Now let us analyse and examine the proposed hypotheses:

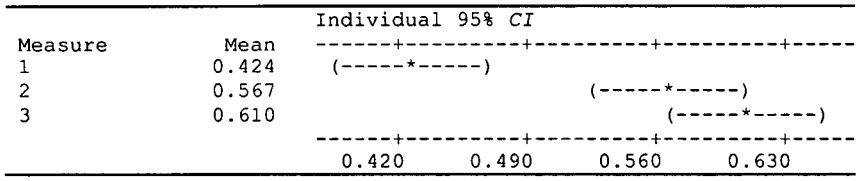
H1: There are no differences in retrieval performance among the three different similarity measures (the angle, the distance, and the distance angle integrated similarity measures), the five different information retrieval models (the angle, distance, conjunction, disjunction, and ellipse retrieval models), and their interactions.

A two-way ANOVA method was used to examine this hypothesis because it involved two factors and they interacted with each other (Glass 1995). (See tables 1, 2, 3, and figure 1.)

Table 1: Result (I) of H1

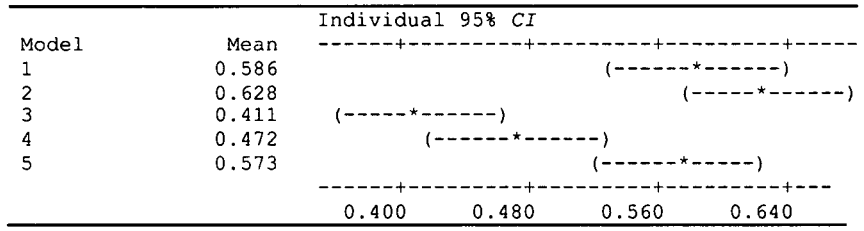
Source	*df*	*SS*	*MS*	*F*	*p*
Measure	2	1.5236	0.7618	20.91	.000
Model	4	1.5370	0.3843	10.55	.000
Interaction	8	0.5011	0.0626	1.72	.095
Error	225	8.1979	0.0364		
Total	239	11.7596			

Table 2: Result (II) of H1



Measure 1 (*ME1*) = angle similarity; Measure 2 (*ME2*) = distance similarity; Measure 3 (*ME3*) = distance angle integrated similarity

Table 3: Result (III) of H1



Model 1 (*MO1*) = angle retrieval; Model 2 (*MO2*) = distance retrieval; Model 3 (*MO3*) = conjunction retrieval; Model 4 (*MO4*) = disjunction retrieval; Model 5 (*MO5*) = ellipse retrieval

Measure 1 (*ME1*) = angle similarity; Measure 2 (*ME2*) = distance similarity; Measure 3 (*ME3*) = distance angle integrated similarity; Model 1 (*MO1*) = angle retrieval; Model 2 (*MO2*) = distance retrieval; Model 3 (*MO3*) = conjunction retrieval; Model 4 (*MO4*) = disjunction retrieval; Model 5 (*MO5*) = ellipse retrieval

In fact, the hypothesis contains three sub-hypotheses: performance among the three similarity measures, performance among the five information retrieval models, and their interactions as well. The analytical data demonstrate that there were significant differences in retrieval performance among the three different similarity measures (p = .000 < α = .05); there were significant differences in retrieval performance among the five different

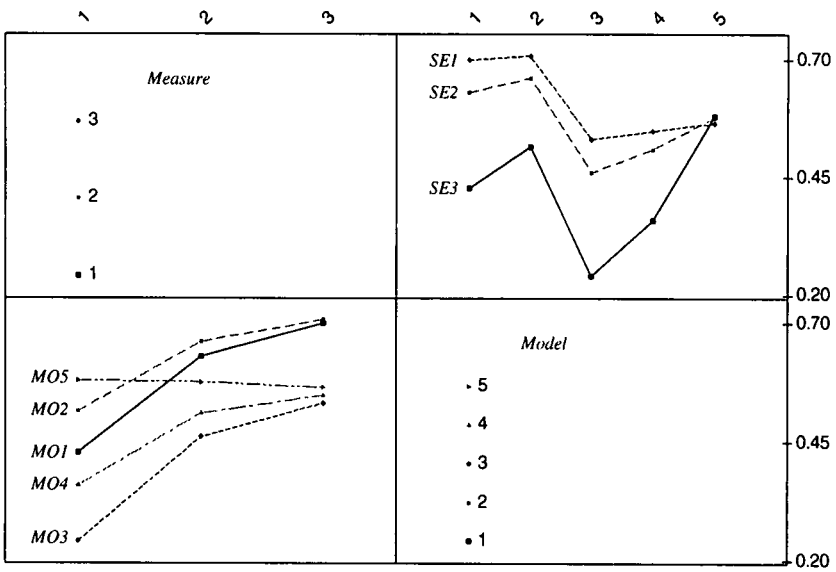


Figure 1: Interaction plot-data means for H1. Measure 1 (ME1) = angle similarity; Measure 2 (ME2) = distance similarity; Measure 3 (ME3) = distance angle integrated similarity; Model 1 (MO1) = angle retrieval; Model 2 (MO2) = distance retrieval; Model 3 (MO3) = conjunction retrieval; Model 4 (MO4) = disjunction retrieval; Model 5 (MO5) = ellipse retrieval

information retrieval models ($p = .000 < \alpha = .05$); and there were no significant differences in retrieval performance among their interactions ($p = .095 > \alpha = .05$).

Since two of these three sub-hypotheses are rejected, further analysis is needed to determine the reasons for hypothesis rejections. A follow-up step was to apply the Tukey's pair-wise comparison technique to determine what factors resulted in these rejections. The comparisons for the three similarity measures and the five information retrieval models are illustrated in tables 12 and 13 respectively.

Tukey's pair-wise comparisons for the three similarity measures

- Family error rate = 0.0500
- Individual error rate = 0.0190
- Critical value = 3.34
- Intervals for (column level mean)—(row level mean)

Table 4: Tukey's pair-wise comparisons among the similarity measures

	Angle *SM1*	Distance *SM2*
Distance *SM2*	−0.2210	
	−0.0658	
Distance angle integrated based *SM3*	−0.2639	−0.1205
	−0.1087	0.0347

1 = angle similarity measure; 2 = distance similarity measure; 3 = distance angle integrated similarity measure

The final results in table 4 provide the endpoints of the confidence intervals for the differences between means corresponding to the levels represented by the row and column headings. The difference between means is considered significant if its confidence interval fails to include zero. In other words, if two numbers in a cell of the table are both positive (negative), then the corresponding mean difference is regarded as significant. This rule is used to determine where a hypothesis is rejected and which factors cause the rejection. Clearly, the pairs (2, 1) and (3, 1) in table 12 meet the condition, suggesting that there were significant differences in performance between the distance similarity measure and the angle similarity measure, and between the distance angle integrated similarity measure and the angle similarity measure. Poor performance of the angle similarity measure accounted for the hypothesis rejection. The curves (the newly developed measure curve at the top, the distance measure curve in the middle, and the angle measure curve at the bottom) in figure 1 confirm these results.

Tukey's pair-wise comparisons for the five information retrieval models

- Family error rate = 0.0500
- Individual error rate = 0.00641
- Critical value = 3.89
- Intervals for (column level mean)—(row level mean)

Using the same approach mentioned earlier, we find that in table 5 the pairs (3, 1), (3, 2), (4, 2), and (5, 3) satisfy the condition. This indicates that there were significant differences in retrieval performance between the conjunction retrieval model and angle retrieval model, between the

Table 5: Tukey's pair-wise comparisons among the retrieval models

	1	2	3	4
2	−0.1589			
	0.0753			
3	0.0580	0.0998		
	0.2922	0.3340		
4	−0.0033	0.0386	−0.1784	
	0.2310	0.2728	0.0559	
5	−0.1043	−0.0625	−0.2794	−0.2181
	0.1299	0.1717	−0.0452	0.0161

1 = angle retrieval model; 2 = distance retrieval model; 3 = conjunction retrieval model; 4 = disjunction retrieval model; 5 = ellipse retrieval model

Table 6: Result of H2

	N	*M*	*SD*	*SE* mean
H2.1	16	0.699	0.147	0.037
H2.2	64	0.588	0.181	0.023

95% CI for mu H2.1—mu H2.2: (0.023, 0.200)
t test mu H2.1 = mu H2.2 (vs >): $t = 2.58$, $p = .0077$, $df = 27$
Note: The axis is the defined retrieval effectiveness

conjunction retrieval model and distance retrieval model, between the disjunction retrieval model and distance retrieval model, as well as between the ellipse retrieval model and the conjunction retrieval model. It is apparent that the poor performance of the conjunction model played a critical role in the rejections. Good performance of the distance retrieval model also made its contribution to the rejection. The curves in figure 1 confirm the findings where the curve of the distance model is located at the top and the curve of the conjunction model is located at the bottom.

H2: The distance angle integrated similarity measure within angle retrieval model achieves better performance than the distance angle integrated similarity measure within all the distance-based retrieval models (see table 6 and figure 2).

H2_1 and *H2_2* stand for the distance angle integrated similarity measure within the cosine (angle) retrieval model and the distance angle integrated similarity measure within all the distance-based retrieval models

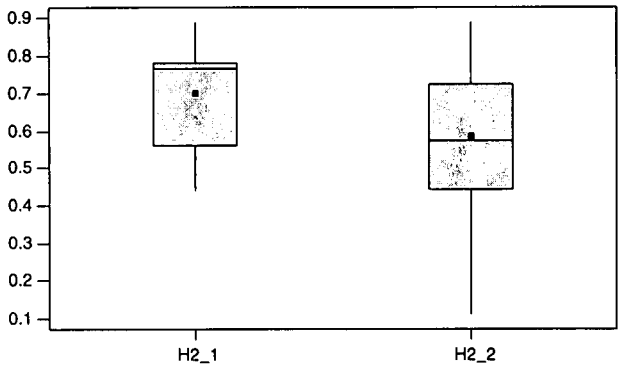


Figure 2: Boxplot H2.1 and H2.2
***Note*: The axis is the defined retrieval effectiveness**

Table 7: Result of H3

	N	*M*	*SD*	*SE* mean
H3.1	32	0.703	0.137	0.024
H3.2	48	0.548	0.179	0.026

95% CI for mu H3.1—mu H3.2: (0.085, 0.226)
t test mu H3.1 = mu H3.2 (vs >): $t = 4.40$, $p = .0000$, $df = 76$

respectively. The data show the hypothesis was accepted because of p (= .0077) < α (= .05).

H3: The distance angle integrated similarity measure within retrieval models based on one reference point achieves better performance than the distance angle integrated similarity measure within all models based on two reference points (see table 7 and figure 3).

H3_1 and *H3_2* stand for the distance angle integrated similarity measure within retrieval models based on one reference point and the distance angle integrated similarity measure within the two-reference-point-based retrieval models respectively. The data show the hypothesis was accepted due to p (= .0000) < α (= .05).

H4a: The distance angle integrated similarity measure achieves better performance than the angle similarity measure within the angle retrieval model (see table 8 and figure 4).

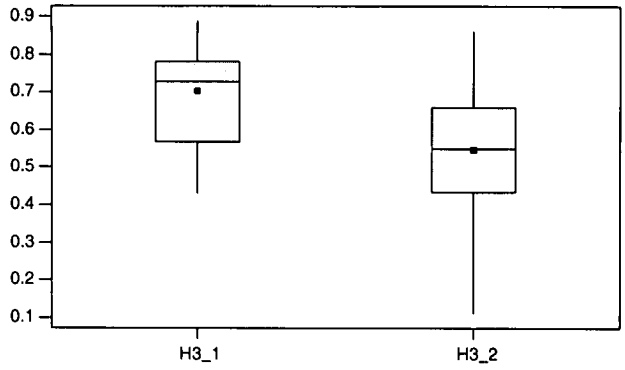


Figure 3: Boxplots of H3.1 and H3.2

Table 8: Result of H4a

	N	*M*	*SD*	*SE* mean
H4.3	16	0.699	0.147	0.037
H4.2	16	0.631	0.184	0.046

95% *CI* for mu H4.3—mu H4.2: (−0.052, 0.189)
t test mu H4.3 = mu H4.2 (vs >): $t = 1.17$, $p = .13$, $df = 28$

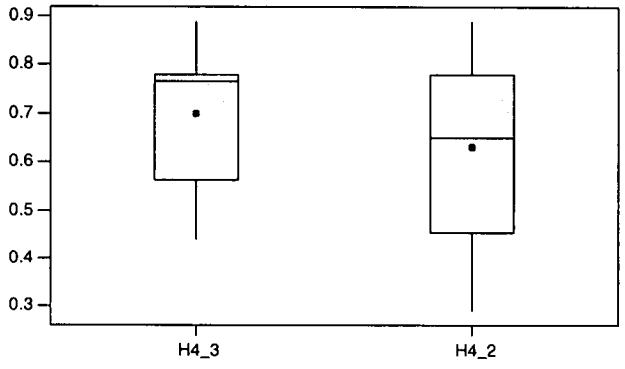


Figure 4: Boxplots of H4.3 and H4.2

H4_3 and *H4_2* stand for the distance angle integrated similarity measure and the angle similarity measure within the cosine retrieval model respectively. The data show that the hypothesis was rejected, since p (= .13) > α (= .05). Observe that, although the distance angle integrated similarity measure did not achieve better retrieval performance than the angle similarity measure within the angle retrieval model as predicted, the mean

Table 9: Result of H4b

	N	M	SD	SE mean
H4.3	16	0.699	0.147	0.037
H4.1	16	0.427	0.171	0.043

95% CI for mu H4.3—mu H4.1: (0.157, 0.387)
t test mu H4.3 = mu H4.1 (vs >): $t = 4.83$, $p = .0000$, $df = 29$

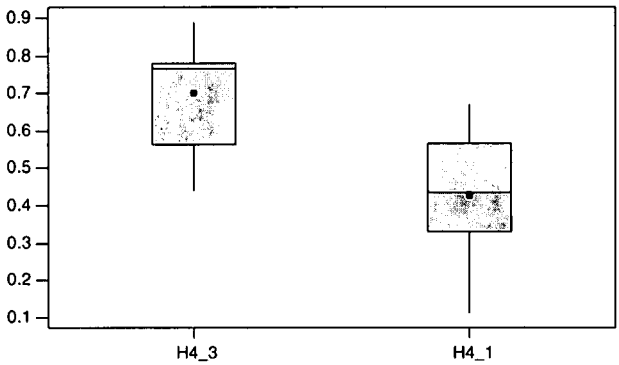


Figure 5: Boxplots of H4.3 and H4.1

of the former (0.699) was still larger than that of the latter (0.631), and the distribution range of the former (0.56, 0.79) was narrower than that of the latter (0.46, 0.78) (see table 8 and figure 4).

H4b: The distance angle integrated similarity measure achieves better performance than the distance similarity measure within the angle retrieval model (see table 9 and figure 5).

H4_3 and *H4_1* stand for the distance angle integrated similarity measure and the distance similarity measure within the angle retrieval model respectively. The data shows that the hypothesis was accepted since p (= .0000) < α (= .05).

H5a: The distance angle integrated similarity measure achieves better performance than the angle similarity measure within the distance-based retrieval models (see table 10 and figure 6).

H5_3 and *H5_2* stand for the distance angle integrated similarity measure and the angle similarity measure respectively within the distance-based retrieval models. The data show the hypothesis was rejected since p (= .15) > α (= .05).

Table 10: Result of H5a

	N	M	SD	SE mean
H5.3	64	0.588	0.181	0.023
H5.2	64	0.551	0.208	0.026

95% CI for mu H5.3—mu H5.2: (−0.032, 0.105)
t test mu H5.3 = mu H5.2 (vs >): $t = 1.06$, $p = .15$, $df = 123$

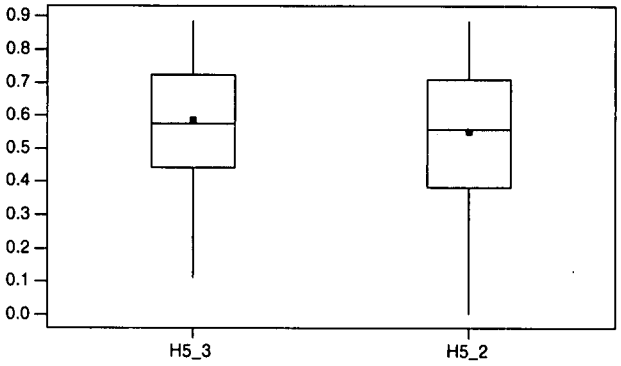


Figure 6: Boxplots of H5.3 and H5.2

H5b: The distance angle integrated similarity measure achieves better performance than the distance similarity measure within the distance-based retrieval models (see table 11 and figure 7).

H5_3 and *H5_1* stand for the distance angle integrated similarity measure and the distance similarity measure respectively within the distance-based retrieval models. The data shows that the hypothesis was accepted since p (= .0000) < α (= .05).

The result of the hypothesis H1 suggests the distance angle integrated similarity measure achieved the best performance among the three tested similarity measures. The results were consistent with the results of the hypotheses H4(b) and H5(b). In addition, although the hypotheses H4a and H5a were rejected, the means of the distance angle integrated similarity measure (0.699 and 0.588) in the three tests were still larger than those of the angle similarity measure in the corresponding categories (0.631 and 0.551), respectively. Furthermore, the standard deviation values in H4a and H5a show performance of the distance angle integrated similarity measure was more stable than that of the angle similarity measure because the standard deviation values of the former were smaller than these of the

Table 11: Result of H5b

	N	Mean	SD	SE mean
H5.3	64	0.588	0.181	0.023
H5.1	64	0.423	0.250	0.031

95% CI for mu H5.3—mu H5.1: (0.089, 0.241)
t test mu H5.3 = mu H5.1 (vs >): $t = 4.28$, $p = .0000$, $df = 114$

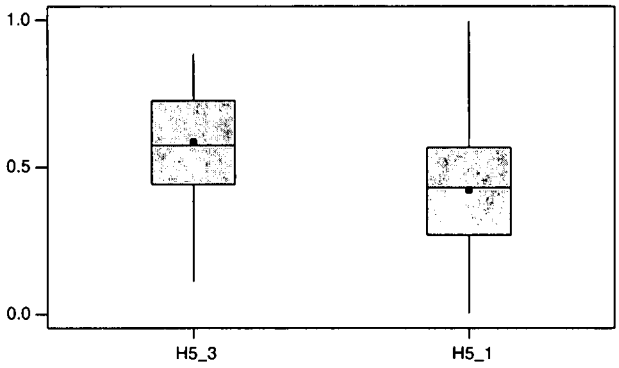


Figure 7: Boxplots of H5.3 and H5.1

latter (0.147 vs. 0.184, and 0.181 vs. 0.208). The Tukey's pair-wise comparison indicates that the distance angle integrated similarity measure achieved better performance than the angle similarity measure, the distance similarity measure achieved better performance than the angle similarity measure, and the conjunction retrieval model achieved the worst retrieval performance among the five information retrieval models. The angle, distance, conjunction, and disjunction information retrieval models worked best in the distance angle integrated similarity measure, second best in the distance similarity measure, and third best in the angle similarity measure. Performance of the ellipse retrieval model was inconsistent with the same pattern. On the contrary, it did best with the angle similarity measure, second best with the distance similarity measure, and third best with the distance angle integrated similarity measure (see figure 1). However, there was no significant difference among the three similarity measures within the ellipse model.

Note that the distance angle integrated measure did not achieve better performance than that of the angle similarity measure within all retrieval models significantly. Poor performance of the ellipse retrieval model within the distance angle integrated similarity measure can explain why the newly

Table 12: Summary of the hypotheses ($\alpha = .05$)

H	SAM	DBRM	ABRM	SM1	SM2	SM3	ORPM	TRPM	p	Result
H2	*t* test	√	√			√			.0077	Accepted
H3	*t* test					√	√	√	.0000	Accepted
H4a	*t* test		√	√		√			.1300	Rejected
H4b	*t* test		√		√	√			.0000	Accepted
H5a	*t* test	√		√		√			.1500	Rejected
H5b	*t* test	√			√	√			.0000	Accepted
H1a	ANOVA			√	√	√			.0000	Accepted
H1b	ANOVA	√	√						.0000	Accepted
H1c	ANOVA	√	√	√	√	√			.0950	Rejected

H = hypothesis, SAM = statistical analysis method, DBRM = distance based retrieval model, ABRM = angle based retrieval model, SM1 = angle similarity measure, SM2 = distance similarity measure, SM# = distance angle integrated similarity measure, ORPM = one-reference-point-based retrieval model, TRPM = two-reference-point-based retrieval model

developed measure did not achieve better performance than that of the angle similarity measure.

The results of the hypotheses H2 and H3 illustrate that the distance angle integrated similarity measure achieved better retrieval performance in the angle-based retrieval model than that of the distance-based retrieval models, and better performance in the models based on one reference point than those based on two reference points. It means that the impacts of the newly developed similarity measure on information retrieval models varied in different information retrieval model types.

For a summary of the hypotheses tests see table 12.

Conclusion

In this experimental study, the information retrieval performance of the distance angle integrated similarity measure was investigated.

The analytical results demonstrate that

- The distance angle integrated similarity measure within the angle retrieval model achieved better performance than the distance angle integrated similarity measure within all the distance-based retrieval models.

- The distance angle integrated similarity measure within the models based on one reference points achieved better performance than the distance angle integrated similarity measure within all retrieval models based on two reference points.
- The distance angle integrated similarity measure did not achieve better performance than the angle similarity measure within the angle retrieval model.
- The distance angle integrated similarity measure achieved better performance than the distance similarity measure within the angle retrieval model.
- The distance angle integrated similarity measure did not achieve better performance than the angle similarity measure within the distance-based retrieval models.
- The distance angle integrated similarity measure achieved better performance than the distance similarity measure within the distance-based retrieval models.
- There were significant differences in retrieval performance among the three different similarity measures (the angle, the distance, and the distance angle integrated similarity measures), the five different information retrieval models (the angle, distance, conjunction, disjunction, and ellipse retrieval models), and their interactions.

The Tukey's pair-wise comparison suggests that both poor performance of the angle similarity measure and poor performance of the conjunction information retrieval model in the experimental study were primarily responsible for the significant differences respectively.

In conclusion, in this study the distance angle integrated similarity measure achieved better retrieval performance than the distance similarity measure. These findings demonstrate the soundness of the distance angle integrated similarity measure from the practical perspective. One exception was that the distance angle integrated measure did not achieve better performance than the angle similarity measure within the five information retrieval models but the mean of the former was still larger than that of the latter.

Notice that in this study two parameters a and c were selected as 1/0.9 and 0.5 respectively. These two parameters a and c influence the result of the newly developed similarity measure, from the distance and angle respectively.

All conclusions were drawn on the basis of the two parameters. Changing either of these two parameters may result in different results.

Notice that the distance angle integrated measure failed to achieve better performance than the angle similarity measure in three different circumstances: within all models, angle model, and distance model. It suggests that in these circumstances the direction (angle) made positive contributions to the retrieval effectiveness. The parameter c (in these cases, $c = 0.5$) played a critical role in the direction (angle) impact on the distance angle integrated measure. If the parameter c increases, it might improve the retrieval effectiveness of the distance angle integrated measure.

There are many similarity measures available. It is widely recognized that each similarity measure has its own strength and weakness. Some similarity measures work better than others under some circumstances. It is quite normal. It is impossible for a similarity measure to outperform others under any circumstances. Identifying under which circumstance a similarity measure works well is always a research topic.

It would be premature to generalize the findings beyond the scope of the study environment. The performance of the similarity measures may vary with the size and coverage of the database, search environment, the two parameters in the distance angle integrated similarity measure, or the number of subjects (queries) size.

Future research directions on the issue include, but are not limited to, the application of this similarity measure to the construction of a visual space for a visual retrieval tool and comparisons between this similarity measure and other similarity measures such as the iso-content-based angle similarity measure (Zhang and Rasmussen 2001).

References

Allan, J., A. Leuski, R. Swan, and D. Byrd. 2001. Evaluation combinations of ranked lists and visualizations of inter-document similarity: Task the TREC-6 Interactive Track. *Information Processing & Management* 37 (3): 435–58.

Atlam, E.S., M. Fuketa, and K. Morita. 2003. Documents similarity measurement using field association terms. *Information Processing & Management* 39 (6): 809–24.

Atlam, E.S., M. Fuketa, K. Morita, and J.I. Aoe. 2000. Similarity measurement using term negative weight and its application to word similarity. *Information Processing & Management* 36 (5): 717–36.

Bartell, B.T., G.W. Cottrell, and R.K. Belew. 1998. Optimizing similarity using multi-query relevance feedback. *Journal of the American Society for Information Science* 49 (8): 742–61.

Burrell, Q.L. 2005. Measuring similarity of concentration between different informetric distributions: Two new approaches. *Journal of the American Society for Information Science and Technology* 56 (7): 704–14.

Calado, P., M. Cristo, and M.A. Goncalves. 2006. Link-based similarity measures for the classification of Web documents. *Journal of the American Society for Information Science and Technology* 57 (2): 208–21.

Chen, Z.X., and B. Fu. 2007. On the complexity of Rocchio's similarity-based relevance feedback algorithm. *Journal of the American Society for Information Science and Technology* 58 (10): 1392–1400.

Croft, W., and D. Harper. 1979. Using probabilistic models of information retrieval without relevance information. *Journal of Documentation* 35:285–95.

Cronin, B. 1994. Tiered citation and measures of document similarity. *Journal of the American Society for Information Science* 45 (7): 537–8.

Egghe, L. 2006. Properties of the n-Overlap Vector and n-Overlap Similarity Theory. *Journal of the American Society for Information Science and Technology* 57 (9): 1165–77.

Egghe, L., and C. Michel. 2003. Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. *Information Processing & Management* 39 (5): 771–807.

Ellis, D., H.J. Turner, and P. Willett. 1993. Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management* 3 (2): 128–49.

Falkowski, B.J. 1998. On certain generalizations of inner product similarity measures. *Journal of the American Society for Information Science* 49 (9): 854–8.

Fricke, M. 1997. Information using likeness measures. *Journal of the American Society for Information Science* 48 (10): 882–92.

Glass, G.V. 1995. *Statistical methods in education and psychology.* Boston: Allyn and Bacon.

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

Hamers, L., Y. Hemeryck, and G. Herweyers. 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing & Management* 25 (3): 315–18.

Kim M.C., and K.S. Choi. 1999. A comparison of collocation-based similarity measures in query expansion. *Information Processing & Management* 35 (1): 19–30.

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

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): 342–351.

McGill, M., M. Koll, and T. Noreault. 1979. *An evaluation of factors affecting document ranking by information retrieval systems.* Syracuse: School of Information Studies, Syracuse University.

Michel, C. 2001. Ordered similarity measures taking into account the rank of documents. *Information Processing & Management* 37 (4): 603–22.

Na, S.H., I.S. Kang, and J.H. Lee. 2007. Adaptive document clustering based on query-based similarity. *Information Processing & Management* 43 (4): 887–901.

Nuchprayyoon, A. 1996. GUIDO: A usability study of its basic information retrieval operations. PhD diss., University of Pittsburgh.

Olsen, K.A., and R.R. Korfhage. 1994. Desktop visualization. In *Proceedings 1994 IEEE Symposium on Visual Languages,* 239–44. St. Louis, MO: IEEE.

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

Qin, J. 2000. Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. *Journal of the American Society for Information Science* 51 (3): 166–80.

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

—. 1985. A theoretical framework for defining similarity measures for Boolean search request formulations, including Some ExperimentalResults. *Information Processing & Management* 21 (6): 501–24.

Robertson, S.E., and K. Sparck Jones. 1976. Relevance weighting of searching terms. *Journal of the American Society for Information Science* 27:129–46.

Robertson, S.E., and S. Walker. 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,* 16–24. Philadelphia, PA: ACM.

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

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

Torvik, V.I., M. Weeber, and D.R. Swanson. 2005. A probabilistic similarity metric for medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology* 56 (2): 140–58.

Travison, D. 1987. Term co-occurrence in cited/citing journal articles as a measure of document similarity. *Information Processing & Management* 23 (3): 183–94.

Tudhope, D., and C. Taylor. 1997. Navigation via similarity: Automatic linking based on semantic closeness. *Information Processing & Management* 33 (2): 233–42.

Watters, C., and H. Wang. 2000. Rating news documents for similarity. *Journal of the American Society for Information Science and Technology* 51 (9): 793–804.

Zhang, J. 2000. A visual information retrieval tool. *In Proceedings of the 63 Annual Meeting of the American Society for Information Science,* 248–57. Medfort, NJ: Information Today.

—. 2001. *TOFIR*: A tool of facilitating information retrieval; Introduce a visual retrieval model. *Information Processing & Management* 37 (4): 639–57.

Zhang, J., and R.R. Korfhage. 1999a. *DARE*: Distance and angle retrieval environment: A tale of the two measures. *Journal of the American Society for Information Science* 50 (9): 779–87.

—. 1999b. A distance and angle similarity measure. *Journal of the American Society for Information Science* 50 (9): 772–8.

Zhang, J., and E.M. Rasmussen. 2001. Developing a new similarity measure from two different perspectives. *Information Processing & Management* 37 (2): 279–94.

—. 2002. An experimental study on the iso-content-based angle similarity measure. *Information Processing & Management* 38 (3): 325–42.

Appendix 1: Query summary

Query no.	Query description
1	Airplane accidents always cause damage. Both military and civil airplanes can crash. The reasons for airplane accidents are mechanical accidents or terrorism, according to reports. Please find some news about air transportation in the AP (1989) database.
2	Weather has a significant impact on human beings. Bad weather, like floods, heavy rain, heavy snow, and storm even threaten human lives. Please find some reports about the impact of severe weather on human beings in the AP (1989) news database.
3	Earthquakes are a main natural catastrophe. They cause casualties. Please find some reports about earthquakes that occurred in the world in the AP (1989) news database.
4	Great political changes took place in the eastern countries. Dubcek and Havel replaced the former leaders of Czechoslovakia and became new leaders. Please search the news about the power shift of Czechoslovakia in the AP (1989) database.
5	Drug abuse and drug traffic is a big social problem. Many crimes are related to drugs (crack, cocaine, etc.). Drug abuse is extremely harmful to children. Please find reports about this issue in the AP (1989) news database.
6	Noriega, the leader of Panama, fled the Vatican embassy in Panama after US troops occupied Panama. It caused some dispute between the US and the Vatican. Please find reports about this issue related to Vatican in the AP (1989) news database.
7	Please find reports about bankruptcy filing, job cutting, business downsizing, business merging, enterprise buying, and bankruptcy in the AP (1989) news database.
8	Ceausescu, the former leader of Romania, was executed. Ceausescu was in power for many years. Please find information about Ceausescu and his family in the AP (1989) news database.
9	Please find the news about movies, movie actors, movie actresses, and related issues in the AP (1989) news database.
10	US troops took military action in Panama, which resulted in its soldiers' deaths. Please search reports on the military action, soldier deaths in the action, and the United Nations' attitude.

Appendix 2: Subject summary

Category	Content	Count
Major	Library science	22
	Information science	10
	Other	0
Gender	Male	16
	Female	16
Education level	Undergraduate	0
	Master's	23
	Doctorate	9

Copyright of Canadian Journal of Information & Library Sciences is the property of University of Toronto Press and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.