A Study of Topic Similarity Measures

Ryen W. White
Department of Computing Science
University of Glasgow
Glasgow, Scotland, G12 8RZ.
ryen@dcs.gla.ac.uk

Joemon M. Jose
Department of Computing Science
University of Glasgow
Glasgow, Scotland, G12 8RZ.
jj@dcs.gla.ac.uk

ABSTRACT
In this poster we describe an investigation of topic similarity measures. We elicit assessments on the similarity of 10 pairs of topic from 76 subjects and use these as a benchmark to assess how well each measure performs. The measures have the potential to form the basis of a predictive technique, for adaptive search systems. The results of our evaluation show that measures based on the level of correlation between topics concords most with general subject perceptions of search topic similarity.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]

General Terms
Design, Human Factors

1. INTRODUCTION
Adaptive search systems work on behalf of users to improve the effectiveness of their search [3]. To operate effectively, such systems must identify both the current search topic and when a search has changed (i.e. moved from one topic to another). Whilst research has focused on identifying search topics it is also important to develop an understanding of which techniques can track changes in these topics.

In this study we evaluate a number of topic similarity measures that could form the basis of a predictive technique to track search topic change. Unlike related studies [2], we evaluate these measures using subjects' evaluations as a baseline. We elicit user assessments on the similarity of 10 TREC topic pairs and for each measure we calculate the level of similarity between these pairs. A strong correlation between measures and subject assessments implies concordance with general subject opinion, and since they operate on subjects' behalf, a suitability for use in adaptive search systems. That is, measures that correlate well with subject assessments can be used to detect when, and by how much, a search has changed in terms of the documents assessed relevant.

2. ESTABLISHING BASELINE SIMILARITY
We use topics 101-150 taken from the TREC initiative and the San Jose Mercury News 1991 collection. This collection has on average 55.6 relevant documents for each topic. As an initial assessment of topic similarity we computed the percentage relevant document overlap for each pair of topics (i.e. the size of their intersection divided by the size of their union). If we consider the combinations of Topics *A* and *B* there were 36 pairs of topics with some degree of overlap between their relevant documents, ranging from .327% (topics 111 & 135) to 21.51% (topics 118 & 119). We choose 10 pairs of topics with varying degrees of overlap, including some with high overlap, some with moderate overlap and some with no overlap at all. The topic pairs chosen and their overlap are shown in (a) and (b) of Table 1.

We used the TREC topic information to construct an overview of each of the topics using the *title*, the topic *summary* and the topic *description*. The two highest ranked *relevant* documents for each topic, chosen using a *tf.idf* weighting scheme using the title field as the query, also formed part of the overview.

Subjects were recruited from the staff, undergraduates and postgraduates at the Department of Computing Science, University of Glasgow. A questionnaire was created containing the topic pairs and placed on a website. All 76 subjects provided similarity assessments on the 10 topic pairs and were encouraged to compare topics carefully. They were shown the overview for each topic in the pair and asked, based on the information available to them, to indicate on a 5-point semantic differential the similarity/dissimilarity of each pair. Responses were automatically logged for later analysis.

We compare each measure's perception of topic similarity with general subject perception of similarity. The mean *subject* similarity for each pair is shown in column (c) of Table 1. We assume this is on an interval scale and a reasonable indicator of general subject perceptions. A good measure should generate similarity assessments that correlate closely with such perceptions.

In the next section we describe the similarity measures used.

3. SIMILARITY MEASURES
In this study we use three classes of similarity measure: *association*, *correlation* and *distance*. In this section we describe each of these classes. However, we must first describe the term distributions used by these classes. These distributions were constructed across the each topic overview using:

$$P(t) = \frac{ntf(t)}{\sum_{t \in T} ntf(t)} \qquad \text{where } ntf(t) = \frac{\log_2(tf(t)+1)}{\log_2 |T|}$$

ntf(*t*) is the normalised term frequency [2] of term *t* in the set of terms *T* taken from the topic overview. This set (i.e. all unique terms from the topic title, summary, description and example relevant documents) is extracted and ranked based on *P*(*t*), the probability that a term *t* is relevant *in that topic*. We divide by the sum of all *ntf*(*t*) to ensure the probabilities sum to one.

We calculate the similarity between each pair of topics using the topic distributions and association, correlation and distance measures. Each measure is based on the intersection between the topic sets (i.e. terms that occur in both topic overview).

Association: *Dice's*, *Jaccard's*, *Cosine* and *Overlap*. The topic overviews are represented by an unweighted set of terms *A* and *B*.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

Table 1. Measures of topic pair similarity (shaded = similar, unshaded = dissimilar)

(a) Topic pair	(b) Overlap	(c) Mean	(d) Mode (%)	Normalised Association				Normalised Correlation			Normalised Distance		
				Dice's	Jaccard's	Cosine	Overlap	Spearman	Kendall	Pearson	Euclidean	L_1(norm)	KL
130 & 145	17.65%	3.04	4 (39.47)	2.67	2.53	2.66	2.81	3.27	3.25	3.28	3.51	3.28	3.73
101 & 102	16.67%	3.99	5 (40.79)	2.79	2.72	2.72	2.40	3.99	3.99	3.99	1.99	2.24	1.32
119 & 120	12.32%	3.09	4 (36.84)	3.99	3.99	3.99	3.99	3.32	3.42	3.29	3.99	3.99	3.99
113 & 117	2.86%	2.78	2 (26.32)	2.39	2.31	2.35	2.18	2.43	2.61	2.53	2.60	2.12	1.73
126 & 135	1.42%	2.80	2 (32.89)	2.21	2.14	2.16	1.94	2.67	2.73	2.40	2.62	1.96	1.83
108 & 111	0.34%	1.71	1 (59.21)	2.75	2.68	2.72	2.58	2.73	2.80	2.85	2.47	2.56	1.57
111 & 135	0.33%	1.32	1 (80.26)	1.81	1.77	1.86	2.03	2.26	2.44	1.32	2.09	2.09	1.71
107 & 110	0.00%	1.57	1 (68.42)	1.93	1.88	1.90	1.80	1.84	1.86	1.92	1.32	1.60	1.46
126 & 147	0.00%	1.63	1 (69.74)	1.32	1.32	1.32	1.32	1.32	1.32	2.24	1.63	1.32	1.77
103 & 119	0.00%	1.87	1 (55.26)	2.18	2.12	2.18	2.17	2.07	2.14	2.05	1.72	1.83	1.68
ρ_{all} with (c)	**.828**	**–**	**–**	**.818**	**.818**	**.794**	**.661**	**.830**	**.830**	**.891**	**.636**	**.636**	**.285**

Correlation: *Pearson product moment correlation* (r), *Spearman's rho* (ρ) and *Kendall's tau-b* (τ). All coefficients measure the relationship between the two topics through the term lists extracted from their topic overview. Pearson's r uses actual values of $P(t)$, whereas ρ and τ are non-parametric and use the rank-order of t (based on $P(t)$). Spearman's ρ accounts for the proportion of variability between *ranks* in the two lists, Kendall's τ represents the difference between the probability that the lists are in the same order versus the probability that the lists are in different orders.

Distance: *Euclidean*, *L_1(norm)* and *Kullback-Leibler*. Each measure represents how disparate the two topics are. Since the Kullback-Leibler measure is not symmetric the order of the comparison matters. In this, as in all measures, topic A is assumed to precede topic B. In the next section we describe the evaluation of these measures against the baseline similarity established in Section 2.

4. EVALUATION OF MEASURES

To determine their effectiveness in predicting topic similarity the measures were compared against subject evaluations across the same topic pairs. Table 1 shows the results obtained from all measures. Each column contains the level of similarity between a topic pair as assessed by the measure.

We use a monotonic transformation to convert the raw scores produced by each measure to scores in the range of the average subject similarity assessments, i.e. [1.32,3.99]. Transformation had no effect on the significance of our results and allows the reader to better compare the performance of measures. Unlike the Association and Correlation measures, a lower normalised Distance measure represents a higher similarity. The normalisation of the measures allows for this and the smallest distances are assigned the highest values (e.g. 3.99).

We use the Spearman rank order correlation ρ_{all} to compare each measure with the subject similarity assessments. The range of ρ_{all} is [-1,1] with a value closer to 1 indicating more correlation. We show ρ_{all} in the last row of Table 1. A higher assessment rating (i.e. closer to 5) indicates more topic similarity.

The high correlation between the relevant document overlap and subject similarity (ρ_{all} = .828, p = .003) implies that the overlap between relevant documents for a pair of TREC topics can be a reasonable indicator of topic similarity.

For each topic pair the most common rating (i.e. the mode) and its percentage of the total number of ratings for that pair is shown in column (d). There is more agreement in the ratings assigned for more dissimilar topic pairs. We divided the topic pairs into two classes based on subject assessment ratings; *similar* (i.e. above or equal 3.0) and *dissimilar* (i.e. below 3.0). These are shown as shaded (similar) and unshaded (dissimilar) in Table 1. For each class we measured inter-subject agreement in assigning a similarity rating to a topic pair using Kendall's *coefficient of concordance*. The results showed a significant level of agreement for dissimilar topics ($p < .001$), but not for similar topics ($p > .05$). This suggests that generally subjects were more able to identify topic dissimilarity than similarity.

5. CONCLUSIONS

We have presented an investigation of topic similarity measures using subject assessments as a baseline. The association and correlation coefficients broadly agree with subject assessments on topic similarity. However, the Pearson's product moment correlates most with general subject perceptions of topic similarity. Correlation is arguably more useful than association, since we can derive statistical significance from it. The results show that the correlation coefficients outperform other measures and concord most with subject opinion.

The correlation coefficients conduct comparisons between pairs of term lists created at successive temporal locations. We can regard A and B as the search topic at two distinct points in time and the degree of topic similarity as synonymous with the topic change between these points. Since correlation coefficients concur most with subject perceptions of search topic change, they may be useful to track such change in adaptive search systems. Systems based on this technique would represent search topics by term lists or probability distributions and monitor topic change through differences in the correlation between them.

6. ACKNOWLEDGMENTS

We are grateful to Diane Kelly and Vassilis Plachouras for their helpful comments on this poster.

7. REFERENCES

[1] Harman, D. (1986) 'An Experimental Study of the Factors Important in Document Ranking'. *Proceedings of the 9th ACM SIGIR Conference*, 186-193.

[2] Lee, L. (1999) 'Measures of Distributional Similarity'. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 25-32.

[3] Maes, P. (1994) 'Agents that Reduce Work and Information Overload'. *Communications of the ACM*, 37(7), 30-40.