An Information-theoretic Measure for Document Similarity *

Javed A. Aslam
Department of Computer Science
Dartmouth College
jaa@cs.dartmouth.edu

Meredith Frost
Department of Computer Science
Dartmouth College
Meredith.Frost@dartmouth.edu

ABSTRACT

Recent work has demonstrated that the assessment of pairwise object similarity can be approached in an axiomatic manner using information theory. We extend this concept specifically to document similarity and test the effectiveness of an information-theoretic measure for pairwise document similarity. We adapt query retrieval to rate the quality of document similarity measures and demonstrate that our proposed information-theoretic measure for document similarity yields statistically significant improvements over other popular measures of similarity.

Categories and Subject Descriptors:

H.3.3 [**Information Search and Retrieval**]: Clustering

General Terms: Theory, Experimentation

Keywords: Similarity measures

1. INTRODUCTION

Measuring pairwise document similarity is quintessential to various tasks in information retrieval, such as clustering and some forms of query retrieval. It is therefore important to calculate similarity as effectively as possible, and some research exists comparing the quality of various similarity measures in some contexts [4].

Dekang Lin [3] has investigated the theoretical basis of similarity, and he derived the general form of an information-theoretic measure for object similarity. Similarity may be viewed as a question of how much information two objects have in common and how much they have in difference. Information theory provides a means for quantifying these intuitive notions, being directly concerned with the mathematical expression of information content.

Based on six axioms for similarity, Lin derived the following general form for pairwise object similarity

$$\text{IT-Sim}(A,B) = \frac{I(\text{common}(A,B))}{I(\text{description}(A,B))}$$

where $I(\text{common}(A,B))$ is the information content associated with the statement describing what A and B have in common and $I(\text{description}(A,B))$ is the information content associated with the statement describing A and B. The information content of a statement x is defined by its self-information $\log(1/\pi(x))$ [2] where $\pi(x)$ is the probability of the statement within the world of the objects in question. For objects which can be described by a set $\mathcal{S}$ of independent features s, Lin derives the following instantiation of this principle:

$$\text{IT-Sim}(A,B) = \frac{2 \cdot \sum_{s \in A \cap B} \log \pi(s)}{\sum_{s \in A} \log \pi(s) + \sum_{s \in B} \log \pi(s)}$$

where $\pi(s)$ is the fraction of objects exhibiting feature s.

We may employ this methodology to assess the pairwise similarity of documents if we assume, to a first approximation, that documents are composed of a set of independent term "features." The probability $\pi(t)$ is simply the fraction of corpus documents containing term t, and we need only generalize the above formulation to account for the fact that "normalized" documents may contain a "fraction" of a feature. For each document d and term t, let $p_{d,t}$ be the fractional occurrence of term t in document d; thus, $\sum_t p_{d,t} = 1$ for all d. Two (normalized) documents A and B share $\min\{p_{A,t}, p_{B,t}\}$ amount of term t in "common," while they contain $p_{A,t}$ and $p_{B,t}$ amount of term t individually. We may then infer the following

$$\text{IT-Sim}(A,B) = \frac{2 \cdot \sum_t \min\{p_{A,t}, p_{B,t}\} \log \pi(t)}{\sum_t p_{A,t} \log \pi(t) + \sum_t p_{B,t} \log \pi(t)}.$$

2. TESTING SIMILARITY MEASURES

We adapt the process of query retrieval in the TREC competition to test the effectiveness of similarity measures. Based on the assumption that relevant documents are more similar to each other than to those that are non-relevant [5], the technique is as follows:[1]

(1) For each document relevant to a query retrieval topic, use each similarity measure to retrieve a ranked list of the most similar documents. In essence, treat this document as if it were a query.

(2) Obtain a measurement of the quality of the ranked lists using the TREC evaluation program.

(3) Average the results for all docs within a query, then for all queries, to yield a final number for each TREC corpus.

*This work partially supported by NSF Career Grant CCR-0093131.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

[1]We employ the Porter stemmer and the SMART stop word list to index our corpora.

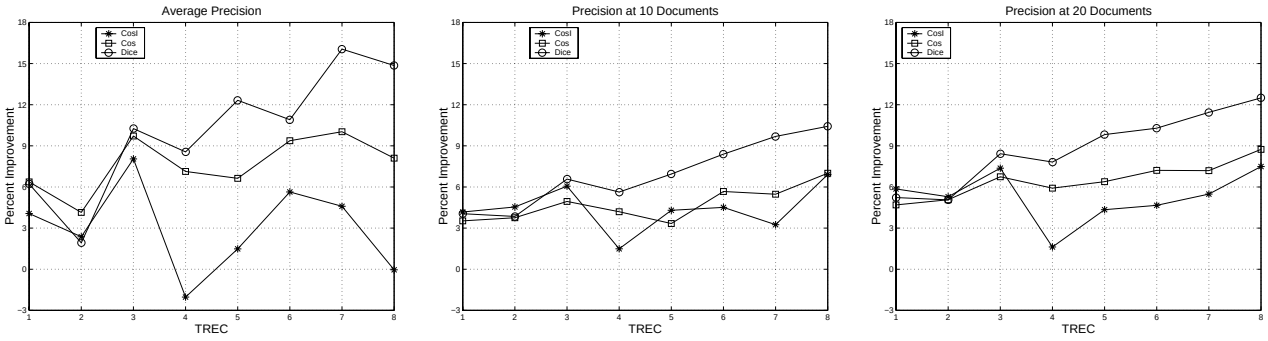

Figure 1: (a) Percentage improvement in mean average precision of information-theoretic similarity (IT) vs. IDF weighted cosine similarity (CosI), unweighted cosine (Cos), and Dice coefficients (Dice). (b) Similar percentage improvements for precision at 10 documents. (c) Similar percentage improvements for precision at 20 documents.

TREC1	IT	CosI	Dice	Cos
IT	X	+4.1	+6.2	+6.4
CosI	−3.9	X	+2.1	+2.2
Dice	−5.9	−2.0	X	+0.1
Cos	−6.0	−2.2	−0.1	X

TREC2	IT	Dice	CosI	Cos
IT	X	+1.9	+2.4	+4.1
Dice	−1.9	X	+0.4	+2.2
CosI	−2.3	−0.4	X	+1.7
Cos	−4.0	−2.1	−1.7	X

TREC3	IT	CosI	Cos	Dice
IT	X	+8.0	+9.7	+10.3
CosI	−7.4	X	+1.5	+2.0
Cos	−8.9	−1.5	X	+0.5
Dice	−9.3	−2.0	−0.5	X

TREC4	CosI	IT	Cos	Dice
CosI	X	+2.1	+9.3	+10.8
IT	−2.0	X	+7.1	+8.6
Cos	−8.5	−6.7	X	+1.3
Dice	−9.8	−7.9	−1.3	X

TREC5	IT	CosI	Cos	Dice
IT	X	+1.5	+6.6	+12.3
CosI	−1.5	X	+5.1	+10.7
Cos	−6.2	−4.8	X	+5.3
Dice	−11.0	−9.6	−5.1	X

TREC6	IT	CosI	Cos	Dice
IT	X	+5.6	+9.4	+10.9
CosI	−5.3	X	+3.5	+5.0
Cos	−8.6	−3.4	X	+1.4
Dice	−9.8	−4.8	−1.4	X

TREC7	IT	CosI	Cos	Dice
IT	X	+4.6	+10.0	+16.1
CosI	−4.4	X	+5.2	+11.0
Cos	−9.1	−4.9	X	+5.5
Dice	−13.8	−9.9	−5.2	X

TREC8	CosI	IT	Cos	Dice
CosI	X	+0.0	+8.2	+14.9
IT	−0.0	X	+8.1	+14.9
Cos	−7.5	−7.5	X	+6.2
Dice	−13.0	−12.9	−5.9	X

Figure 2: Comparing measures of similarity for eight TREC competitions.

3. RESULTS

The graphs shown in Figure 1 show the percentage improvements in average precision, precision at 10 documents, and precision at 20 documents yielded by IT-Sim over a number of popular measures for similarity in our tests conducted using TRECs 1 through 8. We compare IT-Sim against the Dice coefficient (Dice) [3], unweighted Cosine (Cos), and Cosine with an IDF weight (CosI) [1]. Note that in almost every case, IT-Sim out-performs the other measures. In the one case where IT-Sim loses (TREC 4 vs. CosI), the result is relatively close.

The individual corpus tables presented in Figure 2 give a more detailed breakdown of the average precision evaluations. For each corpus, the measures are ranked from best to worst, and the percentage improvement (or decline) of the "row" method vs. the "column" method is given. IT-Sim is clearly the overall winner by winning or nearly tying for first place in every table. In a sign test conducted over all trials, we have shown that in every case where IT-Sim beats any other measure in these tables, the results are statistically significant (well beyond a 95% confidence level).

4. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[3] D. Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, 1998.

[4] M. McGill, M. Koll, and T. Norreault. An evaluation of factors affecting document ranking by information retrieval systems. Technical report, Syracuse University School of Information Studies, 1979.

[5] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.