6-1-1999

Investigating Measures for Pairwise Document Similarity

Jeffrey D. Isaacs
Dartmouth College

Investigating Measures for Pairwise Document Similarity

Jeffrey Isaacs

Advisor: Javed Aslam

Dartmouth College Computer Science Technical Report PCS-TR99-357

Abstract

The need for a more effective similarity measure is growing as a result of the astonishing amount of information being placed online. Most existing similarity measures are defined by empirically derived formulas and cannot easily be extended to new applications. We present a pairwise document similarity measure based on Information Theory, and present corpus dependent and independent applications of this measure. When ranked with existing similarity measures over TREC FBIS data, our corpus dependent information theoretic similarity measure ranked first.

I Introduction

Many similarity measures have been proposed and implemented. Popular examples are the cosine measure, the Okapi measure, and the Dice Coefficient measure. An accurate similarity measure is critical to most information retrieval tasks. Filtering, clustering, and query retrieval are examples that are particularly dependent upon similarity measures.

We have derived an information theoretic similarity measure and three probability models for bag-of-words based document representation. Each probability model makes use of Information Theory, first described in 1948 by Claude Shannon of Bell Telephone Laboratories[1]. Unlike many other empirically derived similarity measures, our measure is derived from basic assumptions about general 'feature' similarity integrated with Shannon's fundamental theorems.

Through experiments over the TREC FBIS database, we tested each probability model application of our similarity measure. The TREC FBIS database consists of a large number of American news media articles indexed into 'qrel' topic groups. We used these topic groups to measure the performance of each similarity measure. Our basic assumption for these tests follows: an accurate similarity measure will always return a higher similarity for two articles in the same 'topic group' than it does for articles not in the same topic group. It should be noted that the TREC topic groups were not created by computer, but rather by government security agency members.

II Derivation

First, we derive an information-theoretic definition of similarity using the following definitions:

- A, B, C — Documents A, B, C, etc.
- I(A,B,C, ...) — Information content of A, B, C (number of bits to encode A, B, C as stated by information theory given a particular probabilistic model)
- $A \cap B$ — 'Features' A and B share in common.

- A Δ B What features A and B have in difference.
  A Δ B = (A - B) U (B - A)

It follows that the information content of documents A and B is the sum of the information content of the features they share and the information content of the features that have in difference. Thus we can fully describe, or encode, A and B by describing their commonalties and their differences:

$$I(A, B) = I(A \cap B) + I(A \Delta B) = I(A \cap B) + I(A - B) + I(B - A)$$

Based on work by Dekang Lin[2], we now propose the following information theoretic definition of similarity for two documents:

$$sim(A,B) = \frac{I(A \cap B)}{I(A,B)} = \frac{I(A \cap B)}{I(A \cap B) + I(A \Delta B)} = \frac{I(A,B) - I(A \Delta B)}{I(A,B)}$$

We now derive three different probability models based on the bag-of-words model. In a bag-of-words representation, documents are described by the words they contain and the frequency of each word in the document. The ordering of the words is unimportant in this model.

We use the following definitions in the derivation of the two corpus dependent probability models:

- Let $D_1, D_2, ..., D_n$ be a set of n documents
- Let $w_1, w_2, ..., w_m$ be a set of words found in these documents

The first corpus dependent probability model(referred to as nats model in data):

- Associated with each document D_i is an m-dimensional probability vector $p_i = (p_{i1}, p_{i2}, ..., p_{im})$ whose entries P_{ij} are the fractional occurrence of word j in doc i.

Thus, $P = (P_{ij})$ is an n x m stochastic matrix.

$$\check{\pi} = (\pi_1, \pi_2, ..., \pi_m)$$

where

$$\pi_j = \frac{1}{n} \sum_{i=1}^{n} P_{ij}$$

- Thus, π_j is the average fractional occurrence of word j in the corpus.

The second corpus dependent probability model(referred to as ‘bin model’ in data):

- Associated with each document D_i is an m-dimensional vector $q_i = (q_{i1}, q_{i2}, ..., q_{im})$ whose entries Q_j is 1 if word j is in doc i, and is 0 if word j is not it doc i.

Thus, $Q = (Q_{ij})$ is an n x m matrix.

$$\check{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$$

where

$$\pi_j = \frac{1}{n}\sum_{i=1}^{n} Q_{ij}$$

- Thus, π_j is the fraction of documents in which word j appears.

We can now choose either probability model and apply it to the definition of similarity obtained earlier:

We consider the similarity of two documents D_r and D_s. By either probability model, D_r contains P_{rj} "amount" of feature/word j and D_s contains P_{rs}. They both share $\min(P_{rj}, P_{rs})$ of that feature in common. They differ by exactly $\max(P_{rj}, P_{rs}) - \min(P_{rj}, P_{rs})$.

As Shannon Information Theory asserts, the 'information' of an event(word) is equal to the negative logarithm of its probability.

Thus,

$$I(D_r \cap D_s) = \sum_{j=1}^{m} \min(P_{rj}, P_{sj}) \cdot (-\log \pi_j)$$

Similarly,

$$I(D_r \Delta D_s) = \sum_{j=1}^{m} (\max(P_{rj}, P_{sj}) - \min(P_{rj}, P_{sj})) \cdot (-\log \pi_j)$$

Therefore,

$$sim(A, B) = \frac{I(A \cap B)}{I(A \cap B) + I(A \Delta B)} = \frac{\sum_{j=1}^{m} \min(P_{rj}, P_{sj}) \cdot (-\log \pi_j)}{\sum_{j=1}^{m} \max(P_{rj}, P_{sj}) \cdot (-\log \pi_j)}$$

Our third application is corpus independent and follows from the above derivation if we restrict the corpus to the two documents being compared:

$$sim(A,B) = \frac{I(A \cap B)}{I(A \cap B) + I(A \Delta B)} = \frac{\sum_{j=1}^{m} \min(P_{rj}, P_{sj}) \cdot (\log \frac{P_{rj} + P_{sj}}{2})}{\sum_{j=1}^{m} \max(P_{rj}, P_{sj}) \cdot (-\log \frac{P_{rj} + P_{sj}}{2})}$$

III Existing Similarity Measures

The Cosine Measure:

The cosine measure can be thought of as similarity being the angle computed by taking the dot product of two bag–of–words vectors:

$$\cos(v_a, v_b) = \frac{v_a \cdot v_b}{\|v_a\|v_b\|} = \frac{v_a \cdot v_b}{\sqrt{(v_a \cdot v_a)(v_b \cdot v_b)}}$$

If we consider such a vector to be a bag-of-words document representation, then:

$$\cos(a,b) = \frac{\sum_{(t \in q) \wedge (t \in d)} f_{d,t} f_{q,t}}{\sqrt{\left(\sum_{t \in q} f_{q,t}^2\right)\left(\sum_{t \in q} f_{d,t}^2\right)}}$$

where $f_{x,t}$ is the frequency of term t in document x.

The Dice Coefficient is defined as follows:

$$dice(a,b) = \frac{\sum_{(t \in q) \wedge (t \in d)} f_{d,t} f_{q,t}}{\left(\sum_{t \in q} f_{q,t}^2\right)\left(\sum_{t \in q} f_{d,t}^2\right)}$$

where $f_{x,t}$ is the frequency of term t in document x.

IV Data

To compare the different similarity measures, we chose the eleven-point average precision test as an evaluation metric. Several definitions will be required before presenting this test. The following contingency table will be helpful:

	RELEVANT	**NON-RELEVANT**

RETRIEVED	$A \cap B$	$\bar{A} \cap B$	B
NOT RETRIEVED	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$\bar{B}$
	A	$\bar{A}$	N

We now define precision and recall:

$$\text{PRECISION} = \frac{|A \cap B|}{|B|}$$

$$\text{RECALL} = \frac{|A \cap B|}{|A|}$$

Informally, precision represents the fraction of documents retrieved by our similarity measure that are 'correct', or deemed relevant. Similarly, recall represents the fraction of the relevant documents which are returned by the similarity measure.

The eleven point average precision(EPAP) test of a *document q with respect to similarity measure* f_{sim} *over a collection of documents S* is computed with the following information for each document d in the collection S:

1. f_{sim} (q, d) [The similarity rating of q and d]
2. relevant(q, d) [A human scoring of 1 if q and d are relevant, 0 otherwise]

With the above information, the test returns 11 pairs of data , for each level of recall and its associated precision:

{{recall-0,precision}, {recall-0.1, precision}, ... {recall-1, precision}}

For example, the second pair represents the precision of the similarity measure f at a recall level of 10%.

Also returned is the eleven-point average, which is the average of the eleven precision values.

<u>Application</u>

We obtained one result for a given similarity measure f_{sim} using the following application of the eleven point average precision:

- For each of approximately twenty(randomly chosen) 'relevant' documents in a given TREC qrel topic(e.g. nuclear non-proliferation), compute the eleven-point precision test of that document with respect to f_{sim} over the TREC FBIS document collection.

- We averaged these twenty sets of data to obtain one set of EPAP results for each of twenty–one TREC topics.

- Moreover, we averaged(See Table 1) the twenty–one TREC average EPAP test results to obtain a final rating for the similarity measure f_{sim}.:

• We ranked the performance of f_{sim}. for each qrel topic, based on the 11-point average precision average of averages. Table 1 contains the average rank over the twenty-one experiments.

Table 1

	IT(nats)	IT(bin)	Dice	Cosine	IT(nocorp)
Recall:					
0%	1.0000	1.0000	1.0000	1.0000	1.0000
10%	0.5680	0.5719	0.5535	0.5389	0.5284
20%	0.3690	0.3774	0.3601	0.3422	0.3314
30%	0.3060	0.3144	0.2893	0.2797	0.2719
40%	0.2713	0.2772	0.2608	0.2504	0.2380
50%	0.2412	0.2467	0.2379	0.2245	0.2096
60%	0.2060	0.2121	0.2131	0.2026	0.1806
70%	0.1867	0.1922	0.1960	0.1873	0.1662
80%	0.1686	0.1722	0.1802	0.1725	0.1553
90%	0.1537	0.1555	0.1616	0.1578	0.1448
100%	0.1369	0.1377	0.1379	0.1379	0.1335
Average precision:	0.3279	0.3325	0.3264	0.3176	0.3054
Rank avg:	2.6667	1.8095	2.4286	3.6190	4.4762

V Conclusion

The information theoretic similarity measure based on the binary corpus–dependent probability model consistently outperformed all other similarity measures we tested. The nats-model performance more closely matched the performance of the Dice measure. Worst performers were the cosine measure and the corpus independent model.

The most significant differences between the similarity measures appear at the early recall stages, i.e. 10%. At this level we see that both of the corpus independent measures surpassed all other measures in precision by 2 percentage points(4% improvement). Towards later stages of recall, all of the measures reached a similar performance limit.

VI Bibliography

1. Shannon, Claude E. and Weaver, Warren, The Mathematical Theory of Communication., Chicago: University of Illinois Press, 1963.

2. Lin, Dekang, "An Information–Theoretic Definition of Similarity.", *Proceedings of the Fifteenth International Conference on Machine Learning* (1998) : pp296–304.