User Judgements of Document Similarity

Mustafa Zengin and Ben Carterette
Department of Computer and Information Sciences
University of Delaware
Newark, DE, USA 19716
{zengin,carteret}@udel.edu

ABSTRACT

Cosine similarity is a term-vector-based measure of similarity that has been used widely in information retrieval research. In this study, we collect user judgments of web document similarity in order to investigate the correlation between cosine similarity and users' perception of similarity on web documents. Experimental results demonstrate that it is hard to deduce that cosine similarity correlates strongly with human judgements of similarity.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]; H.3.4 [**Systems and Software**]: Performance Evaluation

Keywords: document similarity, cosine similarity, users

1. INTRODUCTION

Measures of similarity between documents are widely used in information retrieval: as scores to rank documents, for clustering, for diversity, and more. Most of these measures are based on simple textual features, primarily term counts, document counts, and document lengths. Probably the most widely used is the *cosine similarity*, a measure of the distance between two weighted vectors in a vocabulary space.

Since most uses of such similarity measures are meant to help users with some task, it is worth asking whether they correspond to the notion of similarity that users actually have. The field of IR has always compared query-document similarity measures to human judgements of relevance—this is the foundation of effectiveness evaluation—but there is very little work comparing *document-document* similarity measures to human opinion. In this paper we describe an experiment to collect human judgements of document-document similarity and determine the extent to which cosine similarity captures them.

2. USER EXPERIMENTS

2.1 Experimental Design

The experiments were performed on Amazon Mechanical Turk (AMT) [1], an online crowdsourcing labor marketplace where requesters submit Human Intelligence Tasks (HITs) with some constraints and workers complete the tasks for a fee. Each HIT submitted to workers consisted of a set of instructions about the task, five queries with descriptions that clarify their information need, and five pairs of fully-rendered web pages. Users were asked the following questions: (1) Examine the two web pages shown side by side and judge whether each would help a user achieve the stated description of an information need. (2) Rate the two web pages by how similarly they would help a user achieve the stated description of an information need on a 1–5 scale. Three sets of radio buttons were shown to users to collect the answers. The former question had two answer options: "True" which states the web document helps a user to achieve the stated description of the information need, and "False" otherwise. We provided a rating scale to the users to clarify the similarity ratings for the latter question:

Similarity rating scale

1. Either or both of the two pages do not provide any information for the stated description. Even if two pages are identical, if they are not relevant to the stated description, you should select 1.
2. The two pages give a small amount of similar information for the stated description. They may differ on many points, or one page may be much more substantive than the other, or differ in other ways; they only overlap on a few things.
3. The two pages convey similar information for the stated description. They may differ on some points, or one page may offer more information than the other, but there is some overlap in information relevant to the description.
4. The two pages are mostly equivalent for the stated description; you would be happy to see either one of them in search results. They contain the same information needed to achieve the description, though they may present it in a different order, or along with other information that is not relevant to the description, or with other minor differences.
5. The two pages are essentially equivalent and relevant to the description. They contain exactly the same information, even if they differ in some cosmetic respects (sidebars, top/bottom matter, etc).

Hit properties

In our experiment we set a completion time limit of 3 hours for each HIT and 7 days for each query batch. We asked each HIT to 3 different users and users were paid $0.40 per completed HIT.

Quality control One of the concerns of requesters about crowdsourcing marketplaces such as AMT is low quality work. Due to the high cost of manual data review we used two methods to automatize the quality control process. First we accepted users having following qualifications: Mechanical Turk masters, 95% or higher HIT approval rate, at least 100 HITs of approved work, and a minimum qualification

SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013), August 1, 2013, Dublin, Ireland.

Table 1: Similarity score distribution for document pairs rated by different workers.

C	s1	s2	s3	s4	s5
s1	50	87	30	18	1
s2	87	132	120	69	8
s3	30	120	236	165	17
s4	18	69	165	186	20
s5	1	8	17	20	6

Table 2: Probability distribution of Table 1.

P	s1	s2	s3	s4	s5
s1	0.269	0.209	0.053	0.039	0.019
s2	0.468	0.317	0.211	0.151	0.154
s3	0.161	0.288	0.415	0.360	0.327
s4	0.097	0.166	0.290	0.406	0.385
s5	0.005	0.019	0.03	0.044	0.115

test score of 30 over 36. The qualification test consists of 4 simple questions having the same design layout as the actual task. We also used a "trap" question in each HIT which has a simple, obvious answer; workers that answered it incorrectly were likely to have submitted low-quality work.

2.2 Materials

We randomly selected 10 queries from TREC 2011 Web Track that had been identified as "faceted". For each query we selected 8 web documents that were marked as relevant (i.e., *rel*, *key* or *nav*) by NIST assessors from the top 50 documents of University of Glasgow's adhoc submission `uogTrA45Vm` [2]. From the selected web documents every possible distinct pair is created for each query. The total number of pairs in our experiment set was 280.

3. DATA ANALYSIS

We collected a total of 840 similarity score judgements (3 for each distinct pair) from 45 workers over 2-weeks of experiment duration. We first investigated the users score agreement on document pairs. Table 1 shows counts of document pairs for which one worker gave the rating corresponding to the column label and another gave the rating corresponding to the row label (the counts in this table add up to $280 \times 6 = 1680$, since 3 workers for each pair produces 6 possible comparisons of ratings). Overall agreement, calculated as the sum of the diagonals divided by the total count, is about 36%, which is about on par with measured human agreement about relevance [4, 3].

Table 2 shows the probability of a worker giving the rating corresponding to row given that another worker gave the rating corresponding to the column (columns sum to 1). Having the largest probabilities on the diagonal shows there is some agreement between users in the *score 2*, *score 3* and *score 4* cases. Since we only included documents that are distinct and relevant to the given query, results in *score 1* and *score 5* diagonal were as expected. Note also that the probabilities on the diagonal create the largest sums with probabilities in neighboring cells above or below. This shows even users do not agree on a certain score in 5-scale scoring, their scores are not random.

Figure 1 shows the average cosine similarity (and confidence interval) of document pairs according to given user scores. Note that there are small increases in cosine similarity from *score 2* to *score 3* and *score 4* to *score 5*, but the standard deviations are wide. This suggests that cosine similarity captures something about human judgements, but perhaps not enough that a difference in cosine similarity of as much as 0.4 could be considered meaningful.

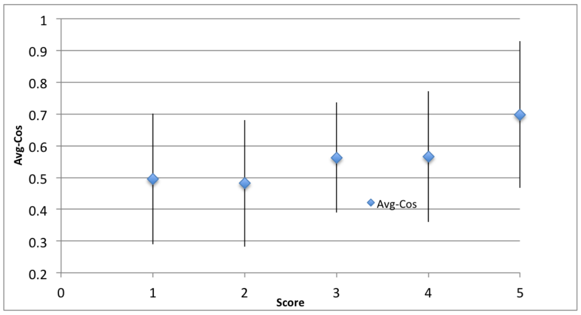

Figure 1: Average cosine similarity of document pairs by worker rating.

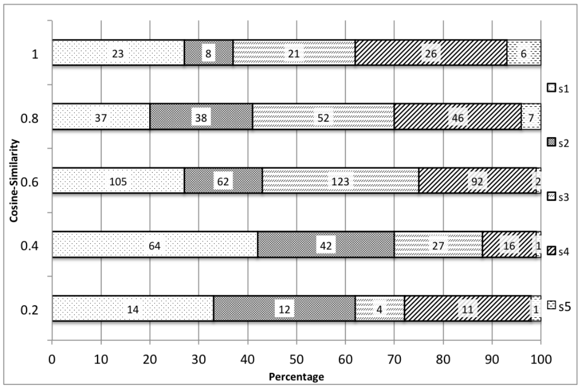

Figure 2: Proportion of ratings in five bins of cosine similarity (labeled on the y-axis by the upper bound of the bin). Numbers on bars are raw counts of document pairs.

Figure 2 shows the percentage of user scores falling into five cosine similarity bins (0.0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1). The share of *score 4* and *score 5* in all bins increases starting from the 0.2–0.4 bin, while the share of *score 1* decreases from the same bin. Nevertheless, it is difficult to say there is a strong association between cosine similarity and given user scores.

The overall correlation between user ratings and cosine similarity is 0.152 by Kendall's τ rank correlation, and 0.189 by linear correlation. While these are significant, they are very low.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1017026. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

4. REFERENCES

[1] Amazon Mechanical Turk. *http://www.mturk.com*

[2] R. McCreadie, C. Macdonald, R. L. T. Santos, I. Ounis. Experiments with Terrier in Crowdsourcing, Microblog, and Web Tracks. *Proc. TREC 2011.*

[3] B. Carterette, P. N. Bennett, D. M. Chickering, S. T.

Dumais. Here or There: Preference Judgements for Relevance *Proc. ECIR 2008*.

[4] E. M. Voorhees Variations in relevance judgments and the measurement of retrieval effectiveness. *Proc. SIGIR 1998*.