Learning User Preferences for Topically Similar Documents

Mustafa Zengin
Department of Computer and Information Sciences
University of Delaware
Delaware, USA
zengin@udel.edu

Ben Carterette
Department of Computer and Information Sciences
University of Delaware
Delaware, USA
carteret@udel.edu

ABSTRACT

Similarity measures have been used widely in information retrieval research. Most research has been done on query-document or document-document similarity without much attention to the user's perception of similarity in the context of the information need. In this study, we collect user preference judgements of web document similarity in order to investigate: (1) the correlation between similarity measures and users' perception of similarity, (2) the correlation between the web document features plus document-query features and users' similarity judgements. We analyze the performance of various similarity methods at predicting user preferences, in both unsupervised and supervised settings. We show that a supervised approach using many features is able to predict user preferences close to the level of agreement between users, and moreover achieve a 15% improvement in AUC over an unsupervised approach.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]; H.3.4 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

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.

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 one of the few works on the subject, Lee et al. studied document similarity measures with human subjects [3]. They compared a number of different similarity measures in terms of their correlation to human judgments of similarity, finding that a model based on Latent Semantic Analysis correlated best to human judgments. However, they only looked at very short news articles outside of the context of any search or information seeking task.

In this paper, we describe an experiment to collect human judgments of document-document similarity for web pages with respect to a web information need, and investigate the extent to which similarity measures capture them. Because asking users to express a similarity judgment between two documents is difficult [6], we present a novel *triplet*-based approach that only requires users to say whether one document is more or less similar to a reference document than another. As we will show, this simplifies the problem such that user agreement about similarity is quite high. We then compare published measures of similarity to our user preferences and show that we can train a classifier that predicts user preferences at a level close to agreement between users.

2. USER EXPERIMENT

Preference judgements have been used in the field of IR in order to help assessors make finer distinctions between relevancy levels of documents [2]. Comparison studies between graded absolute judgements and preference judgements show that the latter can be done faster and with about the same level of agreement [1].

We designed a preference-based experiment to collect human judgements about document *similarity.* In this section we describe the experiment setting, experimental data, and the preferences we collected.

2.1 Queries and Documents

We selected 10 queries from the TREC Web Track 2009—2012, specifically queries that we judged to have a clear and unambiguous topic description. For each query we selected 8 highly relevant and easily readable web documents (based on sentence and document lengths, vocabulary and document style) from the first four pages of Google search results for the query.

2.2 Task Design

We present a participant with *three* highly relevant documents (referred to as a *triplet*) along with a query and the primary topic description from the Web track. One of the

CIKM'15, October 19–23, 2015, Melbourne, VIC, Australia.

ACM 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806617 .

Table 1: User experiment statistics

# queries	10
# documents per query	8
# documents	80
# triplets per task	14
# tasks completed	160
# distinct triplets	1680
# preference judgements collected	2240
# participants	21
# man-hour	33

documents of the triplet is displayed at the top and the two others are displayed below it side by side. We use D_t to denote the document at the top, D_l the document to the left, and D_r the document to the right. The participant is asked to choose which of D_l or D_r is *more similar* to D_t in terms of satisfying the given information need.

Since we have 8 documents per query, there are a total of 168 triplets ($= 8 \cdot \binom{7}{2}$) covering all possible placement of 3 documents for each query. For any given D_t there are 21 possible $\langle D_l, D_r \rangle$ pairs. In order to ensure that each pair would be judged at least once, and that some would be judged twice (so that we could evaluate agreement), we assigned 14 of these to one participant and 14 to another in such a way that guaranteed that all 21 preferences would be judged at least once, and exactly 7 would be judged twice for each D_t in each query.

All participants of the experiment were graduate students. They were paid 8 US dollars per hour of work.

Table 1 shows some statistics of the collected data. All tasks were completed, resulting in 1,680 ($= 10 \cdot 8 \cdot \binom{7}{2}$) distinct triplets judged and 560 ($= 10 \cdot 8 \cdot 7$) that were judged twice.

2.3 Agreement

We investigated agreement among participants that worked on the same document triplets. Overall agreement, calculated as the total number of identical preferences over the total number worked on, is about 71% (402/560), which is above previously-reported human agreement about document relevance [1, 4]. We believe agreement is high because we carefully chose topics and documents of a high quality. Also, it seems that it is easier for participants to judge similarity relative to a reference document (in this case our top document D_t) than to judge similarity on an absolute scale. Zengin and Carterette reported much lower agreement for the latter case [6].

3. SIMILARITY METHODS

We will use the data collected from our participants in an experiment to determine the ability of similarity measures and machine-learned classifiers to capture our participants' notion of similarity. Similarity measures like cosine similarity, Jaccard distance, and others have a long history in IR. Recently, Whissell and Clarke proposed that most similarity measures are composed of three components: a term weighting method, a normalization technique, and a distance measure [5]. The term weighting method determines the importance of a term occurring in the document. Normalization is used to adjust the term weights in order to normalize the effect of document length. Distance measures quantify the distance between two document vectors.

Table 2: Term weighting functions for standard similarity methods.

short name	definition
B	1,if $tf_{it} > 0$,else 0
T	tf_{it}
L	log(1+tf_{it})
TI	tf_{it} log($\frac{n}{n_t}$)
LI	log(1+tf_{it}) log($\frac{n}{n_t}$)
I	log($\frac{n}{n_t}$)
TF	tf_{it} log($\frac{cf_{max}}{cf_t}$)
LF	log(1+tf_{it}) log($\frac{cf_{max}}{cf_t}$)
F	log($\frac{cf_{max}}{cf_t}$)

Table 2, due to Whissell and Clarke, provides various types of term weighting methods. Combining a term weighting method from that table with a normalization scheme (which could be Manhattan (M), Euclidean (N), or none) and a distance measure (Euclidian (E), Jaccard (J), or Cosine (C)) produces a similarity measure. For example, a measure called TJM corresponds to using tf_{il} for term weighting, Jaccard distance, and Manhattan normalization, while a measure called TE corresponds to using tf_{il} for term weighting, Euclidian distance, and no normalization.

Whissell and Clarke also showed how the BM25 scoring function can be used as a similarity measure:

$$OK(D_j, D_i) = \sum_{t \in D_i \cup D_j}^{m} \frac{tf_{jt}(k_1+1)}{tf_{jt} + k_1 b_j} \frac{tf_{it}(k_1+1)}{tf_{it} + k_1 b_i} log\left(\frac{n}{n_t}\right) \tag{1}$$

Here $b_i = (1-b) + b\frac{dl_i}{avgdl}$, and k_1 and b are free parameters. The idf weight at the end could be dropped as well, producing an alternative measure referred to as OKTF.

3.1 Similarity Based on Features

In addition to computing similarity based on terms in documents, we could compute similarity based on other features of the document, or of the two documents, or of the documents and a query. For example, we could compute the similarity between two documents on the basis of the number of query terms that appear in their URLs; the similarity would simply be the absolute difference between the number of query terms in the URL of D_i and the number in the URL of D_j.

Features we used are derived from those in the LETOR [7] datasets and include text-based features such as document length, query term frequency counts in document title and body, query-document score (using standard retrieval scores like BM25, language modeling and cosine similarity in vector space), term counts normalized by length, and web-specific features such as URL length, URL depth (i.e. how deep in a tree of subfolders the page is, as captured by the number of slashes in the URL), and the number of outlinks from a document. We refer to Tables 3 and 4 in Section 4 below for lists of feature classes and specific features.

4. EXPERIMENTS

In this section we analyze the effectiveness of "standard" similarity measures and document/query-document features as described in Section 3 for predicting the similarity preferences of participants from Section 2. We then present a supervised learning method that can use similarity measures, document features, and document-query features to learn user preferences.

4.1 Classification Effectiveness of Individual Features

Let Sim(Q,D_i,D_j) be the similarity between D_i and D_j with respect to query Q. We will define a simple binary classifier *classify* that predicts only a *left* or *right* user preference. Given a similarity method *sim*, the documents in a triplet D_t,D_l,D_r, and the query Q, if the similarity between the top document D_t and left document D_l is greater than the similarity between the top document D_t and right document D_r (i.e. $sim(Q, D_t, D_l) > sim(Q, D_t, D_r)$), then the output of $classify(sim, Q, D_t, D_l, D_r)$ is *left*. Otherwise it is *right*. Note that no training is necessary: the predicted class is based solely on whether similarity between two documents is greater than the same similarity measure between two other documents. We can then compare these predictions to the actual user preferences obtain as described above.

We first calculated the predictions of *classify* with each of the similarity measures derived from Whissell and Clarke's framework along with the BM25 similarity methods. We evaluated predictions using classification accuracy, area under the ROC curve (AUC), and Pearson correlation between predictions.

4.1.1 Standard similarity measures

Figure 1 summarizes the classification performances of similarity measures by AUC. As the figure shows, AUC is determined primarily by the method used for term weighting. Within any given term weighting scheme, there is only a small amount of variation due to distance measure and normalization. Figures for classification accuracy and correlation show very similar results, so they have been excluded. This suggests that term weighting plays the greatest role in in classification performance, and specifically that the B, L, I, and F weightings giving the best predictions of user preference. Results for classification accuracy and correlation are essentially the same as those for AUC, so we have omitted them for space.

Figure 2 shows AUC results for the OK similarity measure with different values of free parameters k and b. The best classification performance is achieved with higher values of b and lower values of k, though increasing b causes a steeper decrease in AUC as k increases. Overall, though, the measure is fairly robust to parameter values. The other BM25-based similarity measure OKTF is not shown here, but it shows a similar pattern with regards to making changes to k and b. However, it is more resistant to performance decrease when increasing k.

4.1.2 Document and query/document features

The *sim* function used in our classifier does not have to be a standard similarity measure; as suggested in Section 3.1 it could be a feature of the document or the document/query pair. In this section we use such features to predict user preferences, again evaluating by accuracy, AUC, and correlation.

Table 3 and Table 4 summarize the classification performances of document and query-document features respectively. Because we tested a large number of features, we only report a select subset; in particular, when a feature can be computed for different document fields (URL, title, body, or full text), we report only the field that gives maximum performance.

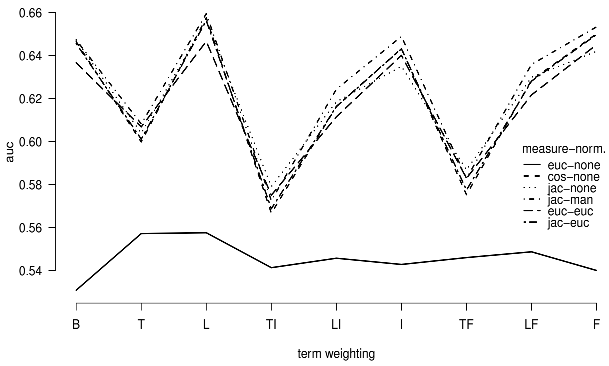

Figure 1: AUC of standard similarity methods

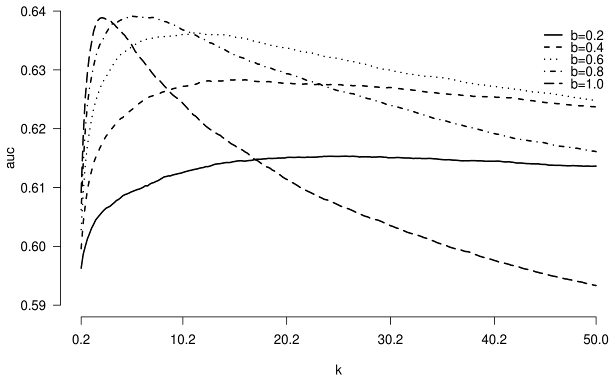

Figure 2: AUC of OK method with k and b values

Table 3: Pearson correlation, accuracy and AUC of document features on document fields. Only document fields with maximum values are reported. † represents URL, ‡ represents title, ⋆ represents body, and ∗ represents full text.

Feature Desc.	Cor.	Acc.	AUC
stream length	0.073∗	0.538∗	0.530∗
number of slash in URL	0.016†	0.512†	0.518†
length of URL	-0.055†	0.477†	0.471†
outlink number	0.075∗	0.536∗	0.523∗

Compared to standard similarity measures in Figures 1 and 2, these features have substantially lower effectiveness for predicting user preferences: AUC ranges between 0.47 and 0.57, only reaching the lower echelons of AUCs shown in those figures. This suggests that overlapping terms in documents make a bigger difference to the user's notion of topical similarity than the overlap of terms between documents and the query.

4.2 Learning User Preference

We next used a supervised method to learn a classifier. We employed a random forest on individual features and combinations of features. For an individual feature such as the OK similarity measure, the feature vector consists

Table 4: Pearson correlation, accuracy and AUC of query-document features on document fields. Only document fields with maximum values are reported. $\dagger$ represents URL, $\ddagger$ represents title, $\star$ represents body, and $\ast$ represents full text. tn represents term number, tf represents term frequency, tr represents term ratio, sl represents stream length.

Feature Desc.	Cor.	Acc.	AUC
covered query tn	-0.220$\dagger$	0.504$\ast$	0.503$\dagger$
covered query tr	-0.025$\dagger$	0.504$\ast$	0.502$\dagger$
sum of tf	0.038$\dagger$	0.524$\dagger$	0.540$\dagger$
min of tf	-0.008$\dagger$	0.508$\dagger$	0.502$\dagger$
max of tf	0.057$\ast$	0.528$\ast$	0.538$\dagger$
mean of tf	0.033$\dagger$	0.522$\dagger$	0.538$\dagger$
var. of tf	0.058$\ddagger$	0.533$\ddagger$	0.534$\dagger$
sum of sl norm.tf	0.098$\ast$	0.546$\ast$	0.543$\ast$
min of sl norm.tf	0.017$\ast$	0.508$\ast$	0.521$\ast$
max of sl norm.tf	0.094$\ast$	0.548$\ast$	0.555$\ast$
mean of sl norm.tf	0.101$\star$	0.552$\star$	0.537$\star$
variance of sl norm.tf	0.145$\dagger$	0.562$\dagger$	0.568$\dagger$
sum of tf*idf	0.043$\dagger$	0.526$\dagger$	0.542$\dagger$
min of tf*idf	0.015$\dagger$	0.515$\dagger$	0.514$\dagger$
max of tf*idf	0.036$\dagger$	0.519$\dagger$	0.524$\ast$
mean of tf*idf	0.046$\dagger$	0.528$\dagger$	0.543$\dagger$
variance of tf*idf	0.035$\dagger$	0.523$\dagger$	0.540$\dagger$
boolean model	-0.018$\dagger$	0.508$\star$	0.510$\dagger$
vector space model	0.044$\ast$	0.522$\ast$	0.534$\ast$
bm25	0.078$\ast$	0.538$\ast$	0.540$\ast$
lm-dir	0.047$\ast$	0.523$\ast$	0.536$\ast$
lm-jm	0.059$\ast$	0.530$\ast$	0.538$\ast$

of $sim(Q, D_t, D_l)$ and $sim(Q, D_t, D_r)$, and the true class is the user preference for D_l or D_r (relabeled to 0 or 1). For combinations of features, we use the same idea, with two similarity calculations for each feature in the model.

Table 5 and Figure 3 summarize the performance of models using OK and OKTF as individual features, a model using all document and query/document features, a model using all standard similarity measures as features, and finally a model using all of the above.

The AUC of the OK measure with $k = 8.2$ and $b = 0.6$ is 0.639 in the unsupervised setting (see Fig. 2). Training with OK increases its classification performance by 3.3%. Using all document and query/document features in combination gives an AUC of 0.719, 27% higher than any individual feature in that set and 9% higher than OK on its own. Using all standard similarity measures together gives an AUC of 0.739, 12% higher than OK but only 3% higher than the document + query/document features. Using all of the above features together further increases AUC by a small amount, and produces classifiers which have accuracies close to the agreement between participants reported in Section 2.3 (67.8% vs 71%).

These results suggest that both classes of features do quite a good job of capturing user preferences—almost to the extent that users can predict each others' preferences.

Table 5: Supervised learning with individual and combination of features

Feature Desc.	Cor.	Acc.	AUC
OKTF (k=8.2, b=0.6)	0.205	0.603	0.652
OK (k=8.2, b=0.6)	0.214	0.608	0.660
Document+Query/Doc Features	0.296	0.649	0.719
Standard Sim. Methods	0.344	0.672	0.739
All	0.354	0.678	0.749

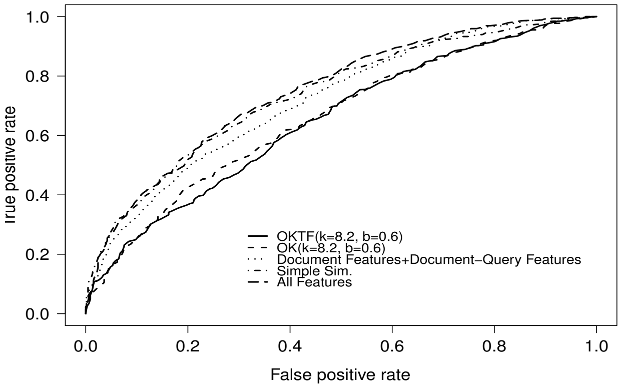

Figure 3: ROC curve of supervised models

5. CONCLUSIONS

We have presented an experiment on having users judge similarity between documents by expressing a binary preference for one document being more similar to a reference document than another. We compared those preferences to standard tf/idf-based similarity measures, unsupervised feature-based similarity, and a supervised classifier including both. We found that the latter achieves accuracy close to the level of human agreement (which in turn is high for an IR task). We next intend to investigate tasks for which this classifier could be useful—tasks that are helped by clustering is a clear direction for future work.

6. REFERENCES

[1] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: preference judgments for relevance In *Proceedings of the ECIR*, pages 16-27, 2008

[2] M. E. Rorvig. The simple scalability of documents. *JASIS*, 41(8):590-598, 1990

[3] Michael D. Lee and Matthew Welsh. An empirical evaluation of models of text document similarity. In *CogSci2005*, 2005.

[4] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697-716, September 2000.

[5] John S. Whissell and Charles L. A. Clarke. Effective measures for inter-document similarity. In *Proceedings of the 22nd ACM international conference on Information and knowledge management*, 2013

[6] M. Zengin and B. Carterette. User judgements of document similarity. *Proceedings of the SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013)* ,2013

[7] Qin, Tao, Tie-Yan Liu, Jun Xu, and Hang Li. LETOR: A benchmark collection for research on learning to rank for information retrieval. In *Information Retrieval 13, no. 4* (2010): 346-374.