Effective Measures for Inter-Document Similarity

John S. Whissell Charles L.A. Clarke
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
{jswhisse,claclark}@uwaterloo.ca

ABSTRACT

While supervised learning-to-rank algorithms have largely supplanted unsupervised query-document similarity measures for search, the exploration of query-document measures by many researchers over many years produced insights that might be exploited in other domains. For example, the BM25 measure substantially and consistently outperforms cosine across many tested environments, and potentially provides retrieval effectiveness approaching that of the best learning-to-rank methods over equivalent features sets. Other measures based on language modeling and divergence from randomness can outperform BM25 in some circumstances. Despite this evidence, cosine remains the prevalent method for determining inter-document similarity for clustering and other applications. However, recent research demonstrates that BM25 terms weights can significantly improve clustering. In this work, we extend that result, presenting and evaluating novel inter-document similarity measures based on BM25, language modeling, and divergence from randomness. In our first experiment we analyze the accuracy of nearest neighborhoods when using our measures. In our second experiment, we analyze using clustering algorithms in conjunction with our measures. Our novel symmetric BM25 and language modeling similarity measures outperform alternative measures in both experiments. This outcome strongly recommends the adoption of these measures, replacing cosine similarity in future work.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering—*similarity measures, algorithms*

General Terms

Experimentation

Keywords

Clustering; Similarity Measures

CIKM'13, October 27 - November 01 2013, San Francisco, CA, USA

http://dx.doi.org/10.1145/2505515.2505526.

1. INTRODUCTION

In recent years, supervised learning-to-rank algorithms have largely replaced unsupervised query-document measures for search [4, 5, 10, 26]. Despite this shift, there is a wealth of information from years of research into unsupervised query-document measures that can be leveraged to enhance results in none-search domains. An example is the use of BM25 [13] as an inter-document similarity measure, such as those used in clustering, as well as many other applications.

Document clustering is a heavily study area of data mining research [2, 7, 8, 14, 15, 16, 20, 21, 24]. In much of this research cosine with tf-idf term weighting remains the defacto standard inter-document similarity measure, as well as in many other domains. This fact is surprising, given that it is known that it performs very poorly in the search domain relative to other measures, such as BM25 [13]. BM25, on the other hand, is competitive with supervised learning-to-rank algorithms in some situations [18]. It might therefore be a more reasonable basis for an inter-document similarity measure. Recent research results corroborate this idea [20]. Alternatively, one could also use other unsupervised query-document measures — known to outperform cosine with tf-idf term weighting for search — and expect similarly improved results.

To explore this idea, we present a number of novel inter-document similarity measures based on BM25 [13], language modeling [22], and divergence from randomness (*DFR*) [1]. We provide experiments, including one in the clustering domain, demonstrating the effectiveness of several of our novel inter-document similarity measures. Our experiments indicate that our novel symmetric BM25 and language modeling inter-document similarity measures are the most effective of those we consider in this paper, both for clustering and for more general tasks. This suggests that these measures should replace cosine with tf-idf term weighting in all future work.

The remainder of this paper proceeds as follows. In Section 2, we discuss preliminaries. In Section 3 we present our novel BM25, language modeling, and DFR based inter-document similarity measures. In Section 4, we show that the nearest neighbors of documents, when using several of our novel measures, are more likely to share the same labels than when using competing measures. This result suggests that they more accurately capture how humans categorize documents. In Section 5, we show similarly positive results with respect to actual clusterings. This result is demonstrated using a number of clustering algorithms and well-

studied datasets. For both of our experiments, our symmetric BM25 and language modeling inter-document similarity measures out-perform the others. We consider parameter sensitivity in our experiments, showing that our best performing methods that require parameters have reasonable universal parameterizations. Finally, we provide a concluding discussion in Section 6.

2. PRELIMINARIES

This section provides preliminaries for the rest of the paper, starting with some notation: Let X be a set of documents, and let X_i be the ith document of X. Let m be the number of distinct terms in X. We assume each X_i is represented using the standard "bag-of-words" vector space model representation:

$$X_i = [x_{i1}, x_{i2}, \ldots, x_{im}], \tag{1}$$

where x_{il} is the weight of term l to document i.

When using the above formula to measure inter-document similarity, the x_{il} weights are typically based on two values: 1) tf_{il}, the *term frequency* of l in i, and 2) idf_l, the *inverse document frequency* of l. A standard document clustering definition for idf_l, as well as for other domains, is [20, 25]:

$$\mathrm{idf}_l = \log(\frac{n}{n_l}), \tag{2}$$

where n is the number of documents in X, and n_l is the number of documents in X that contain term l. Less commonly, some form of *collection frequency* is used as part of an x_{il} weight [12]:

$$\mathrm{cf}_l = \sum_{X_i \in X} \mathrm{tf}_{il}. \tag{3}$$

The x_{il} weights may be length normalized. This is most often achieved through Euclidean length normalization:

$$x_{il} = \frac{x_{il}}{\sqrt{\sum_{l=1}^{m} x_{il}^2}}, \tag{4}$$

but Manhattan length normalization may also be used:

$$x_{il} = \frac{x_{il}}{\sum_{l=1}^{m} |x_{il}|}. \tag{5}$$

A substantial fraction of document vector representations used in inter-document similarity measures are formed using some combination of the above formulae, including the nearly ubiquitous tf-idf term weighted vectors:

$$x_{il} = \mathrm{tf}_{il} \log(\frac{n}{n_l}), \tag{6}$$

which are often Euclidean length normalized.

Using vector representations, the inter-document similarity between a pair of documents X_i and X_j can be computed with any number of measures. The most common is cosine:

$$\mathrm{sim}_{\cos}(X_i, X_j) = \frac{X_i \cdot X_j}{||X_i||_2 ||X_j||_2}, \tag{7}$$

where $||X_i||_2$ is the Euclidean length of X_i:

$$||X_i||_2 = \sqrt{\sum_{l=1}^{m} x_{il}^2}. \tag{8}$$

Note that this formula makes cosine invariant to Euclidean length normalization. Alternatives to Eq. 7 include extended Jaccard similarity [17]:

$$\mathrm{sim}_{\mathrm{jac}}(X_i, X_j) = \frac{X_i \cdot X_j}{||X_i||_2^2 + ||X_j||_2^2 - X_i \cdot X_j}, \tag{9}$$

inverse Euclidean distance:

$$\mathrm{sim}_{\mathrm{euc}}(X_i, X_j) = \frac{1}{d_{\mathrm{euc}}(X_i, X_j)}, \tag{10}$$

where $\mathrm{d}_{\mathrm{euc}}(X_i, X_j)$ is the Euclidean distance between X_i and X_j:

$$\mathrm{d}_{\mathrm{euc}}(X_i, X_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - x_{jl})^2}, \tag{11}$$

and the inverse of the Jensen-Shannon divergence if the vectors can be treated as probability distributions:

$$\mathrm{sim}_{\mathrm{js}}(X_i, X_j) = \frac{1}{\frac{1}{2}D(X_i||M) + \frac{1}{2}D(X_j||M)}, \tag{12}$$

where $D(X_i||M)$ is the Kullback-Leibler divergence between X_i and M, the average of the probability distributions X_i and X_j:

$$D(X_i||M) = \sum_{l=1}^{m} x_{il} \log(\frac{x_{il}}{\frac{x_{il}+x_{jl}}{2}}). \tag{13}$$

For all the measures discussed above higher values indicate higher similarity. As noted previously, the most widely used inter-document similarity measure is cosine (Eq. 7) with tf-idf representations (Eq. 6).

3. RANKING FUNCTIONS BASED DOCUMENT SIMILARITY MEASURES

In this section we present our novel inter-document similarity measures. We begin with our BM25 based measures, followed by our language modeling based measures, and finally we present our DFR based measures.

3.1 BM25 Similarity Measures

Since its introduction in the early 1990s, the BM25 formula [13] has been widely adopted, and it has repeatedly proved its value as a ranking function across a large variety of search domains. If Q is a query consisting of terms, then X_i's BM25 score with respect to Q is:

$$\mathrm{S}(Q, X_i) = \sum_{l \in Q} \frac{q_l(k_3+1)}{q_l + k_3} \frac{\mathrm{tf}_{il}(k_1+1)}{\mathrm{tf}_{il} + k_1 \mathrm{bl}_i} \log(\frac{n}{n_l}), \tag{14}$$

where:

$$\mathrm{bl}_i = (1-b) + b\frac{\mathrm{dl}_i}{\mathrm{avgdl}}, \tag{15}$$

q_l is the term frequency of l in the Q, dl_i is the count of the tokens in document X_i:

$$\mathrm{dl}_i = \sum_{l=1}^{m} \mathrm{tf}_{il}, \tag{16}$$

and avgdl is the average dl for documents in X. The values b, k_1, and k_3 are tuning parameters, with $k_1 \geq 0$, $k_3 \geq 0$, and $0 \leq b \leq 1$. Typical values are k_1 and k_3 between 1.2

and 2, and $b = 0.7$. The $\frac{q_l(k_3+1)}{q_l+k_3}$ component of Eq. 14 is often omitted, except when queries are long.

Previous work [9] peripherally suggests that BM25 should be effective in document clustering, but the idea was not carefully explored. More recently, we performed an in-depth study of document clustering term weighting strategies [20], in which we used BM25 term weights of:

$$x_{il} = \frac{\mathrm{tf}_{il}(k_1+1)}{\mathrm{tf}_{il}+k_1\mathrm{bl}_i}, \tag{17}$$

and:

$$x_{il} = \frac{\mathrm{tf}_{il}(k_1+1)}{\mathrm{tf}_{il}+k_1\mathrm{bl}_i}\log(\frac{n}{n_l}), \tag{18}$$

in combination with cosine as inter-document similarity measures. We had positive results using fixed parameter values of $k_1 = 20$ and $b = 1.0$. In this paper, we implement BM25 directly as a general inter-document similarity measure (i.e., without cosine or any other secondary measure). Previous research has shown that cosine is a poor ranking function relative to BM25 [6], we therefore had a strong rationale for believing that our purely BM25 measure would outperform one synthesizing BM25 and cosine. Our experimental results in later sections support this idea.

To use BM25 in an inter-document similarity measure directly we perform the following adjustments to Eq. 14: 1) We replaced Q with a second document X_j, and 2) We replaced the $\frac{q_l(k_3+1)}{q_l+k_3}$ component with $\frac{\mathrm{tf}_{jl}(k_1+1)}{\mathrm{tf}_{jl}+k_1\mathrm{bl}_j}$, requiring that the b used in each component be equal. The first adjustment is just a notational change. The second adjustment ensures symmetry ($\mathrm{S}(X_i, X_j) = \mathrm{S}(X_j, X_i)$) by enforcing $k_3 = k_1$ and equating b. With the above adjustments, Eq. 14 becomes:

$$\mathrm{OK}(X_j, X_i) = \sum_{l=1}^{m} \frac{\mathrm{tf}_{jl}(k_1+1)}{\mathrm{tf}_{jl}+k_1\mathrm{bl}_j}\frac{\mathrm{tf}_{il}(k_1+1)}{\mathrm{tf}_{il}+k_1\mathrm{bl}_i}\log(\frac{n}{n_l}). \tag{19}$$

This is our first inter-document similarity measure using BM25. Considering that an idf component may not always be beneficial [20], we also implemented a second, tf-only, version:

$$\mathrm{OKTF}(X_j, X_i) = \sum_{l=1}^{m} \frac{\mathrm{tf}_{jl}(k_1+1)}{\mathrm{tf}_{jl}+k_1\mathrm{bl}_j}\frac{\mathrm{tf}_{il}(k_1+1)}{\mathrm{tf}_{il}+k_1\mathrm{bl}_i} \tag{20}$$

This is our second inter-document similarity measures using BM25.

It is worth noting that for any term l to have a non-zero contribution to OK and OKTF, it must have $\mathrm{tf}_{il} > 0$ and $\mathrm{tf}_{jl} > 0$. This property is preserved from Eq. 14. In contrast, a previous use of BM25 in an inter-document similarity measure [20] did not have this property. Indeed, none of other inter-document similarity measures that we discuss in this paper have this property, making a strong distinction between OK/OKTF and other measures.

In Section 4 and 5, we will show that OK is a highly effective inter-document similarity measure, performing as well as our novel language modeling approaches, and better than all the other alternatives. We show its b parameter may be fixed at 1, regardless of k_1, with a minimal loss in performance, and that wide range of k_1 values provide reasonable performance. With respect to OKTF, we show it is approximately as effective as the previous use of BM25 with no idf component and cosine [20].

3.2 Language Modeling Similarity Measures

Language modeling ranking functions usually take a fixed query Q and rank X_is by their estimated $p(X_i|Q)$s, i.e., the probability that their language model generated Q. Using Bayes' rule, $p(X_i|Q)$ can be written as:

$$p(X_i|Q) = \frac{p(Q|X_i)p(X_i)}{p(Q)}. \tag{21}$$

Combining the assumption that $p(X_i)$ is uniform [22] with the fact that Q is fixed gives:

$$p(X_i|Q) \propto p(Q|X_i) \tag{22}$$

Thus $\log(p(Q|X_i))$ estimates can be (and typically are) used to rank by document relevance against a fixed query in language modeling ranking functions. Similarly, we use such estimates in our new general language modeling inter-document similarity measure:

$$\begin{aligned}\mathrm{LM}(X_i, X_j) =& \log(p(X_j|X_i)) - \log(p(X_j|X)) \\ &+ \log(p(X_i|X_j)) - \log(p(X_i|X)),\end{aligned} \tag{23}$$

where X_i and X_j are two documents, and X is the entire dataset treated as a single document. Before computing this measure, all documents are Manhattan length normalized to account for varying document lengths.

Specific versions of Eq. 23 are obtained by translating their $\log(p(Q|X_i))$ estimates into a form applicable to a pair of documents, then substituting that form for each of the logs in the equation. LM is a symmetrizing of language model ranking functions that accounts for varying document length and similarity by random chance, where we model chance using $\log(p(*|X))$. Being based on language model ranking functions, which have a strong theoretical backing, makes LM intuitively reasonable as an inter-document similarity measure.

Many of the commonly discussed $\log(p(Q|X_i))$ estimates in the language modeling literature are amenable to operating on document vectors of Manhattan length normalized term counts, and can thus easily be used with LM. In this paper, we consider using two such estimations: 1) a Dirichlet smoothed estimation and 2) a Jelinek-Mercer smoothed estimation, both of which can found in Zhai and Lafferty [22]. We chose these because they are well-known and have been shown to provide good performance in a large number of ranking experiments.

The Dirichlet smoothed estimation of $\log(p(Q|X_i))$ from Zhai and Lafferty [22] may be rewritten as:

$$\begin{aligned}\log(p(X_j|X_i)) =& \sum_{l=1}^{m} x_{jl}\log(1+\frac{x_{il}}{\mu p(l|X)}) \\ &+ \mathrm{dl}_j\log(\frac{\mu}{\mathrm{dl}_i+\mu}) \\ &+ \sum_{l=1}^{m} x_{jl}\log(p(l|X)),\end{aligned} \tag{24}$$

where X_i and X_j are two document vectors, $p(l|X)$ is the collection probability of term l:

$$p(l|X) = \frac{\mathrm{cf}_l}{\sum_{l=1}^{m}\mathrm{cf}_l}, \tag{25}$$

and $x_{il} = \mathrm{tf}_{il}$ (in our case, this is Manhattan length normalized). μ is the smoothing parameter of the function, with $\mu \geq 0$.

A similar rewriting of the Jelinek-Mercer smoothed estimation from Zhai and Lafferty yields:

$$\log(p(X_j|X_i)) = \sum_{l=1}^{m} x_{jl} \log(1 + \frac{\frac{(1-\lambda)x_{il}}{\mathrm{dl}_i}}{\lambda p(l|X)}) + \mathrm{dl}_j \log(\lambda) + \sum_{l=1}^{m} x_{jl} \log(p(l|X)). \quad (26)$$

λ is the smoothing parameter, $0 \leq \lambda \leq 1$. For details on the base forms of Eq. 24 and 26, and more details on language modeling ranking functions in general, readers can consult Zhai and Lafferty [22]. In this paper we will denote the use of Eq. 24 in LM as DLM, and Eq. 26 in LM as JMLM.

We will show in our experiments that both our language modeling inter-document similarity measures are competitive with OK, and out perform all other measures in this paper. As well, we will show that they function well with a wide range of parameter settings. JMLM will be shown to be particularly robust with respect to λ variation.

3.3 DFR Similarity Measures

DFR models [1] provide a method of assigning term weights to documents based on a comparison of their within-document frequency and their collection frequency (Eq. 3). The first step in DFR term weighting is usually to apply length normalization to raw term counts (as opposed to after term weighting, as is typical in other domains such as document clustering). The standards for this are Manhattan (Eq. 5) and a log normalization of the form:

$$x_{il} = \mathrm{tf}_{il} \log(c\frac{\mathrm{avgdl}}{\mathrm{dl}_i}), \quad (27)$$

where c is some positive constant (1 is a common value). After normalization, the final DFR term weight is typically assigned as:

$$x_{il} = P_R(l, X_i) P_M(l, X_i). \quad (28)$$

P_R estimates the risk in assessing l's relevance to X_i. P_M computes the log of the probability of seeing tf_{il} occurrences of l in X_i under a specific model of randomness. There are various methods of computing P_R and P_M. Table 1 lists the ones we consider in this work, for others readers can consult Amati and van Rijsbergen [1].

Table 1: Our DFR term weighting components.

$P_M(l, X_i)$	Short	Equation
Binomial approximation	P	$-\log\binom{\mathrm{cf}_l}{\mathrm{tf}_{il}}(\frac{1}{n})^{\mathrm{tf}_{il}}(\frac{n-1}{n})^{\mathrm{cf}_l - \mathrm{tf}_{il}}$
Bose-Einstein approximation	G	$-\log((\frac{1}{1+\frac{cf_l}{n}})(\frac{\frac{cf_l}{n}}{1+\frac{cf_l}{n}})^{\mathrm{tf}_{il}})$
tf-Expected idf	$I(n_e)$	$-\mathrm{tf}_{il}\log(\frac{n+1}{n_l+0.5})$

$P_R(l, X_i)$	Short	Equation
Ratio of two binomials	B	$\frac{\mathrm{cf}_l+1}{n_l(\mathrm{tf}_{il}+1)}$
Laplace	L	$\frac{1}{\mathrm{tf}_{il}+1}$

By summing over query terms, Eq. 28 produces the ranking function:

$$\mathrm{S}(Q, X_i) = \sum_{l \in Q} x_{il}. \quad (29)$$

The above is known to be competitive with the BM25 ranking function (Eq. 14) when using a variety of P_R and P_M functions [1]. Given this experience, it makes sense that one could also use DFR for term weighting in inter-document similarity measures. In the same vein, we combined Eq. 28 term weightings with cosine, Jaccard, and inverse Euclidean distance to produce inter-document similarity measures.

We crossed all the P_Ms and P_Rs from Table 1 with four length normalizations: none, Manhattan (Eq. 5), Euclidean (Eq. 4), and log (Eq. 27 with $c = 1$). Length normalizations were applied before Eq. 28, as is standard in DFR. Each weighting was used with cosine, Jaccard, and inverse Euclidean distance to produce a total of 72 (2x3x4x3) distinct DFR similarity measures. We denote a specific DFR inter-document similarity measure as: DFR-*A*-*B*-*C*-*D*, where *A* is the P_M used, *B* is the P_R used, *C* is the length normalization, and *D* is the measure.

Our experiments will show that a variety of our DFR methods are effective, with the best out-performed only by our BM25 and language modeling similarity measures. This result, combined with the fact that DFR methods require no tuning parameters, makes them appealing as inter-document similarity measures.

4. NEAREST NEIGHBOR EXPERIMENT

In order to evaluate our inter-document similarity measures, we conducted two experiments. Our nearest neighbor experiment is detailed in this section. The next section discusses our clustering experiment.

The idea behind our nearest neighbor experiment is as follows: *The better a measure is for a given domain, the better it will reflect expert human notions of similarity in that domain.* To evaluate how well our inter-document similarity measures performed in this respect, we selected eight datasets that have been used in many previous document clustering experiments [2, 7, 16, 20, 21, 23, 24, 25]. Table 2 summarizes their properties[1]. These datasets are not clustering specific. The fbis, k1b, tr31, tr41, and tr45 datasets are from TREC[2], the wap dataset is from the Web-Ace project [3], and the re0 and re1 datasets are taken from the Reuters-21578 text categorization text collection 1.0[3]. For more details on these datasets, readers should consult Zhao and Karypis [25].

For each dataset, for $r = 1$ to 100, we computed the r-nearest neighbors for each document of the dataset when using our inter-document similarity measures. We then computed the average fraction of r-nearest neighbors per document for that dataset sharing the same label as that document for each r value. Finally, we averaged the by-dataset results to obtain a single accuracy value for each r. The better our measures are performing, the higher their accuracies. With respect to parameter ranges tested, for OK and OKTF, we performed this test using a range of $k_1 = 0$ to 50

[1] All of these are available at: `http://glaros.dtc.umn.edu/gkhome/views/cluto/download`

[2] `http://trec.nist.gov`

[3] `http://www.research.att.com/~lewis`

Table 2: The datasets used in our experiments.

Data.	Source	# Docs	# Terms	# Classes
fbis	TREC	2463	2000	17
k1b	TREC	2304	13879	6
re0	Reuters	1504	2886	13
re1	Reuters	1657	3758	25
tr31	TREC	927	10128	7
tr41	TREC	878	7454	10
tr45	TREC	690	8261	10
wap	Web-Ace	1560	8460	20

in increments of 2, combined with a range of $b = 0$ to 1 in increments of 0.1. For DLM, we tested using μ values of 0.1, 0.2, 0.4, 0.8, 1.6, etc., up to 6553.6. For JMLM, we tested using $\lambda = 0.1$ to 0.9 in increments of 0.1.

We performed an analysis identical to the above using a wide variety of inter-document similarity measures. Table 3 provides details. Among the measures are the nearly ubiquitous cosine with tf-idf term weighting (TIC), and a use of cosine with BM25 term weights known to outperform it (OKC) [20]. For OKC and K1C, we investigated the same parameter ranges as we did for OK and OKTF.

Table 4 presents the top 10 measures, by average accuracy from $r = 1$ to $r = 100$, from our nearest neighbor experiment. TIC is included in the table for comparison. Only the single best global parameter setting for each measure is reported in the table, as that is our estimation of the performance one can expect in practice (without by-dataset tuning).

The key result to take from Table 4 is the significance hierarchy of the best performers:

$$\begin{aligned} \text{OK} \quad > \quad & \{\text{JMLM}, \text{DLM}\} > \\ & \{\text{DFR-LOG-In-B-COS,DFR-LOG-In-B-JAC,} \\ & \text{OKC,DFR-In-L-JAC}\}. \end{aligned}$$

All of the members of this hierarchy, except OKC, were designed for this paper, and all produce significantly and substantially better nearest neighborhoods than tf-idf with cosine. This outcome strongly suggests that any of them would make a better inter-document similarity measure than tf-idf with cosine, independent of the application. Note that parameter effects on this hierarchy will be discussed below. We will show that it holds for a range of settings for each method, and that parameter selection for individual methods is straightforward.

From Table 4, it is clear that OK performed the best. It was significantly better than all the other top inter-document similarity measures, and might be viewed as the best inter-document similarity of those we consider, at least from an application-independent standpoint.

OK was significantly and substantially better than OKC, corroborating our suggestion in Section 2 that it would be better to use BM25 directly in an inter-document similarity measure instead of combining it with cosine. For the BM25 measures we tested that had no idf component, OKTF did not perform significantly different from K1C (0.671 versus 0.672 average accuracy, with individual results always within 1%). Both were significantly worse than TIC (.676).

With respect to parameters, we found that OK exhibited uniformly strong by-dataset performance with its optimal parameter setting, being at the worse 2.5% behind the highest accuracy method for any r value on any dataset. Fixing $b = 1.0$ in OK produced excellent results. At most, this resulted in a 2% degradation in accuracy relative to the optimal b for a particular k_1 value (see Fig. 4). Although OK performed the best when $b = 1.0$ and $k_1 = 8$, it was still significantly superior to all the other methods when $4 \leq k_1 \leq 12$ and $b = 1.0$. OKTF's best parameter settings were similar to OK.

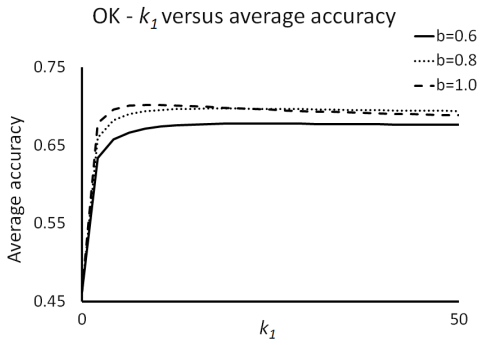

Figure 1: OK's average accuracy as k_1 and b vary.

Both our language modeling inter-document similarity measures performed well, obtaining the second and third highest average accuracies respectively (.691 and .691). Their results were not significantly different from each other, but were significantly better than all the other top measures except OK.

Fig. 4 shows the parameter sensitivity of our language modeling measures with respect to average accuracy. It is notable from the figure that JMLM behaves almost identically from $\lambda = 0.4$ to 0.7, with no statistically significant differences in that range. This suggests that any λ in that range is a reasonable universal parameter setting. For DLM, it is apparent that larger values of μ are harmful to results, with a reasonable range being 0.8 to 3.2. Note that these μ values are substantially smaller than those performing well in DLM's corresponding ranking function because we length normalize documents before computing DLM, whereas the ranking function version does not.

Many of our DFR inter-document similarity measures had robust performance in our nearest neighbor experiment. Six of the top 10 methods by average accuracy were DFR measures, with all P_M and P_Rs being used at least once in those six. DFR-In-B-LOG-COS, DFR-In-B-LOG-JAC, and DFR-In-L-EUC-JAC performed significantly better than any of the other DFR methods, but were inferior to OK and our language modeling measures.

DFR measures that did not use length normalization, or used inverse Euclidean distance similarity, exhibited poor performance in the experiment. It is worth noting that P_R seemed to dictate the optimal normalization. When B was used, log length normalization was best, whereas when L was used, Euclidean normalization was best. The difference from using a non-optimal normalization was often large.

Considering just the inter-document similarity measures in Table 3, OKC performed the best. It was significantly higher than the nearest competitor, TIJM (.682 versus .677). TIJM, on the other hand, was a member of a group of methods with no statistical difference between them: TIJM, TIJ (.676), LIJM (.676), TIEN (.676), TIC (.676), and TIJN (.676). It is noteworthy that TIC was among these methods.

Table 3: Alternative inter-document similarity measures. $\mathbf{cf}_{max}$ is the maximum collection frequency for terms in the dataset. The rest of the notation is as detailed in the body of the paper. For all cases in this table, term weighting is applied first, then length normalization, then the measure.

Short Name	x_{il}	Length Norm.	Measure
BE	1, if $\mathrm{tf}_{il} > 0$, else 0	None	$\mathrm{sim}_{\mathrm{euc}}$
TE	tf_{il}	None	$\mathrm{sim}_{\mathrm{euc}}$
LE	$\log(1 + \mathrm{tf}_{il})$	None	$\mathrm{sim}_{\mathrm{euc}}$
TIE	$\mathrm{tf}_{il} \log(\frac{n}{n_l})$	None	$\mathrm{sim}_{\mathrm{euc}}$
LIE	$\log(1 + \mathrm{tf}_{il}) \log(\frac{n}{n_l})$	None	$\mathrm{sim}_{\mathrm{euc}}$
IE	$\log(\frac{n}{n_l})$	None	$\mathrm{sim}_{\mathrm{euc}}$
TFE	$\mathrm{tf}_{il} \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	None	$\mathrm{sim}_{\mathrm{euc}}$
LFE	$\log(1 + \mathrm{tf}_{il}) \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	None	$\mathrm{sim}_{\mathrm{euc}}$
FE	$\log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	None	$\mathrm{sim}_{\mathrm{euc}}$
BJM	1, if $\mathrm{tf}_{il} > 0$, else 0	Manhattan	$\mathrm{sim}_{\mathrm{jac}}$
TJM	tf_{il}	Manhattan	$\mathrm{sim}_{\mathrm{jac}}$
LJM	$\log(1 + \mathrm{tf}_{il})$	Manhattan	$\mathrm{sim}_{\mathrm{jac}}$
TIJM	$\mathrm{tf}_{il} \log(\frac{n}{n_l})$	Manhattan	$\mathrm{sim}_{\mathrm{jac}}$
LIJM	$\log(1 + \mathrm{tf}_{il}) \log(\frac{n}{n_l})$	Manhattan	$\mathrm{sim}_{\mathrm{jac}}$
IJM	$\log(\frac{n}{n_l})$	Manhattan	$\mathrm{sim}_{\mathrm{jac}}$
LFJM	$\mathrm{tf}_{il} \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	Manhattan	$\mathrm{sim}_{\mathrm{jac}}$
LFJM	$\log(1 + \mathrm{tf}_{il}) \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	Manhattan	$\mathrm{sim}_{\mathrm{jac}}$
FJM	$\log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	Manhattan	$\mathrm{sim}_{\mathrm{jac}}$
BC	1, if $\mathrm{tf}_{il} > 0$, else 0	None	$\mathrm{sim}_{\mathrm{cos}}$
TC	tf_{il}	None	$\mathrm{sim}_{\mathrm{cos}}$
LC	$\log(1 + \mathrm{tf}_{il})$	None	$\mathrm{sim}_{\mathrm{cos}}$
TIC	$\mathrm{tf}_{il} \log(\frac{n}{n_l})$	None	$\mathrm{sim}_{\mathrm{cos}}$
LIC	$\log(1 + \mathrm{tf}_{il}) \log(\frac{n}{n_l})$	None	$\mathrm{sim}_{\mathrm{cos}}$
IC	$\log(\frac{n}{n_l})$	None	$\mathrm{sim}_{\mathrm{cos}}$
TFC	$\mathrm{tf}_{il} \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	None	$\mathrm{sim}_{\mathrm{cos}}$
LFC	$\log(1 + \mathrm{tf}_{il}) \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	None	$\mathrm{sim}_{\mathrm{cos}}$
FC	$\log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	None	$\mathrm{sim}_{\mathrm{cos}}$
OKC	$\frac{\mathrm{tf}_{il}(k_1+1)}{\mathrm{tf}_{il}+k_1 \mathrm{bl}_i} \log(\frac{n}{n_l})$	Euclidean	$\mathrm{sim}_{\mathrm{cos}}$
K1C	$\frac{\mathrm{tf}_{il}(k_1+1)}{\mathrm{tf}_{il}+k_1 \mathrm{bl}_i}$	Euclidean	$\mathrm{sim}_{\mathrm{cos}}$
BEN	1, if $\mathrm{tf}_{il} > 0$, else 0	Euclidean	$\mathrm{sim}_{\mathrm{euc}}$
TEN	tf_{il}	Euclidean	$\mathrm{sim}_{\mathrm{euc}}$
LEN	$\log(1 + \mathrm{tf}_{il})$	Euclidean	$\mathrm{sim}_{\mathrm{euc}}$
TIEN	$\mathrm{tf}_{il} \log(\frac{n}{n_l})$	Euclidean	$\mathrm{sim}_{\mathrm{euc}}$
LIEN	$\log(1 + \mathrm{tf}_{il}) \log(\frac{n}{n_l})$	Euclidean	$\mathrm{sim}_{\mathrm{euc}}$
IEN	$\log(\frac{n}{n_l})$	Euclidean	$\mathrm{sim}_{\mathrm{euc}}$
TFEN	$\mathrm{tf}_{il} \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	Euclidean	$\mathrm{sim}_{\mathrm{euc}}$
LFEN	$\log(1 + \mathrm{tf}_{il}) \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	Euclidean	$\mathrm{sim}_{\mathrm{euc}}$
FEN	$\log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	Euclidean	$\mathrm{sim}_{\mathrm{euc}}$
BJN	1, if $\mathrm{tf}_{il} > 0$, else 0	Euclidean	$\mathrm{sim}_{\mathrm{jac}}$
TJN	tf_{il}	Euclidean	$\mathrm{sim}_{\mathrm{jac}}$
LJN	$\log(1 + \mathrm{tf}_{il})$	Euclidean	$\mathrm{sim}_{\mathrm{jac}}$
TIJN	$\mathrm{tf}_{il} \log(\frac{n}{n_l})$	Euclidean	$\mathrm{sim}_{\mathrm{jac}}$
LIJM	$\log(1 + \mathrm{tf}_{il}) \log(\frac{n}{n_l})$	Euclidean	$\mathrm{sim}_{\mathrm{jac}}$
IJN	$\log(\frac{n}{n_l})$	Euclidean	$\mathrm{sim}_{\mathrm{jac}}$
TFJN	$\mathrm{tf}_{il} \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	Euclidean	$\mathrm{sim}_{\mathrm{jac}}$
LFJN	$\log(1 + \mathrm{tf}_{il}) \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	Euclidean	$\mathrm{sim}_{\mathrm{jac}}$
FJN	$\log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	Euclidean	$\mathrm{sim}_{\mathrm{jac}}$
BJ	1, if $\mathrm{tf}_{il} > 0$, else 0	None	$\mathrm{sim}_{\mathrm{jac}}$
TJ	tf_{il}	None	$\mathrm{sim}_{\mathrm{jac}}$
LJ	$\log(1 + \mathrm{tf}_{il})$	None	$\mathrm{sim}_{\mathrm{jac}}$
TIJ	$\mathrm{tf}_{il} \log(\frac{n}{n_l})$	None	$\mathrm{sim}_{\mathrm{jac}}$
LIJ	$\log(1 + \mathrm{tf}_{il}) \log(\frac{n}{n_l})$	None	$\mathrm{sim}_{\mathrm{jac}}$
IJ	$\log(\frac{n}{n_l})$	None	$\mathrm{sim}_{\mathrm{jac}}$
TFJ	$\mathrm{tf}_{il} \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	None	$\mathrm{sim}_{\mathrm{jac}}$
LFJ	$\log(1 + \mathrm{tf}_{il}) \log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	None	$\mathrm{sim}_{\mathrm{jac}}$
FJ	$\log(\frac{\mathrm{cf}_{max}}{\mathrm{cf}_l})$	None	$\mathrm{sim}_{\mathrm{jac}}$
TJS	tf_{il}	Manhattan	$\mathrm{sim}_{\mathrm{js}}$

Our nearest neighbor experiment from this section was not clustering specific. As such, its key result strongly suggests that its best performers (OKC, DLM, and JMLM) should replace tf-idf with cosine as default application independent inter-document similarity measures. In the following section, we will present a document clustering specific experiment that corroborates this key result, and others discussed in this section, for the domain of document clustering.

5. CLUSTERING EXPERIMENT

For our clustering experiment, we used the same datasets, inter-document similarity measures, and parameter ranges as our nearest neighbor experiment. We drew 10 samples of each dataset, of size equal to half the dataset, and computed a document similarity matrix for each sample using each inter-document similarity measure/parameter setting. These matrices were clustered using four algorithms: UPGMA, complete-linkage, direct e1, and agglomerative i2. We selected these four as they are well-known and tested in document clustering [2, 7, 16, 20, 23, 24, 25]. For each clustering algorithm, we obtained its clustering of each matrix with five, 20, 40, and 80 clusters.

We then computed the correspondence each clustering had with its sample labeling using clustering purity (PQ [23]), clustering entropy (EQ [20]) adjusted mutual information (AMI [19]), and the Rand index (RI [11]). In order to perform a similarity-method focused analysis, we further normalized each correspondence measure result by the maximum that correspondence measure obtained on the same parameters, except when using any similarity method. This procedure was similar to that used by Zhao and Karypis [23] to aggregate correspondence measures from clusterings on different datasets with different numbers of clusters. To avoid confusion with the base correspondence measures, we refer to the normalized correspondence measures as *relative*.

As a final step we obtained the combined average of the four relative correspondences for each inter-document similarity measure over its 1280 clusterings (8 datasets x 10 samples x 4 clustering algorithms x 4 number of clusters). The higher the combined average for a particular measure, the better it was performing as its clustering tended to better reflect true labelings.

Table 5 presents the top 10 measures by combined average. As in our nearest neighbor experiment, we only report the best performing parameterizations in the table.

Table 4: The top 10 inter-document similarity measures from our nearest neighbor experiment, as well as TIC, by average accuracy from $r = 1$ to $r = 100$. Significance of the differences between the average accuracies from a Tukey's test with $p = 0.05$ is included. O indicates that the row is significantly higher than the column, X indicates the opposite, and $-$ indicates no statistical difference. For DLM, μ is based on Manhattan length normalized documents.

Rank	Method	Avg. Acc.	Significance Comparison by Rank										
			1	2	3	4	5	6	7	8	9	10	17
1	OK($k_1 = 8, b = 1.0$)	.702	-	O	O	O	O	O	O	O	O	O	O
2	JMLM($\lambda = 0.6$)	.691	X	-	-	O	O	O	O	O	O	O	O
3	DLM($\mu = 1.6$)	.691	X	-	-	O	O	O	O	O	O	O	O
4	DFR-In-B-LOG-COS	.686	X	X	X	-	-	-	-	O	O	O	O
5	DFR-In-B-LOG-JAC	.685	X	X	X	-	-	-	-	-	O	O	O
6	OKC($k_1 = 20, b = 1.0$)	.682	X	X	X	-	-	-	-	-	-	O	O
7	DFR-In-L-EUC-JAC	.681	X	X	X	-	-	-	-	-	-	-	O
8	DFR-G-L-EUC-JAC	.681	X	X	X	X	-	-	-	-	-	-	O
9	DFR-In-L-EUC-COS	.680	X	X	X	X	X	-	-	-	-	-	O
10	DFR-P-B-LOG-JAC	.680	X	X	X	X	X	X	-	-	-	-	O
17	TIC	.676	X	X	X	X	X	X	X	X	X	X	-

Table 5: The top 10 inter-document similarity measures from our clustering experiment by combined average of the relative correspondence measures. Significance of the differences between combined averages from a Tukey's test with $p = 0.05$ is included. O indicates that the row is significantly higher than the column, X indicates the opposite, and $-$ indicates no statistical difference. For DLM, μ is based on Manhattan length normalized documents.

Rank	Method	Rel. Averages					Significance Comparison by Rank									
		Comb.	AMI	RI	EQ	PQ	1	2	3	4	5	6	7	8	9	10
1	DLM($\mu = 12.8$)	.922	.869	.956	.929	.935	-	O	O	O	O	O	O	O	O	O
3	JMLM($\lambda = 0.7$)	.917	.867	.951	.925	.926	X	-	O	O	O	O	O	O	O	O
2	OK($k_1 = 14, b = 1.0$)	.916	.869	.948	.926	.923	X	X	-	O	O	O	O	O	O	O
4	OKC($k_1 = 22, b = 1.0$)	.909	.857	.941	.920	.917	X	X	X	-	O	O	O	O	O	O
5	DFR-In-L-EUC-COS	.907	.856	.940	.918	.915	X	X	X	X	-	O	O	O	O	O
6	DFR-In-L-EUC-JAC	.902	.846	.938	.913	.910	X	X	X	X	X	-	O	O	O	O
7	DFR-In-L-MAN-COS	.901	.847	.938	.912	.907	X	X	X	X	X	X	-	O	O	O
8	TIEN	.899	.842	.940	.909	.906	X	X	X	X	X	X	X	-	O	O
9	LIEN	.899	.840	.938	.910	.906	X	X	X	X	X	X	X	X	-	O
10	TIC	.897	.841	.935	.908	.903	X	X	X	X	X	X	X	X	X	-

Similar to the results in Table 4, those in Table 5 indicate that several of our novel inter-document similarity measures are superior to the alternatives. The best performers in Table 5 have a significance hierarchy of:

$$\text{DLM} > \text{JMLM} > \text{OK} > \text{OKC} > \text{DFR-In-L-EUC-COS}.$$

Again, OKC is the only method in the hierarchy not designed for this paper. One can see that the top three members of the clustering hierarchy, DLM, JMLM, and OK, are identical to those from the previous experiment. Further, when considering each individual correspondence measure, those three still take up the top three ranks (although their order relative to each other changes). This indicates that DLM, JMLM, and OK are robust inter-document similarity measures for clustering. They can and should be adopted as the default inter-document similarity measures in clustering.

OK performed 3rd best overall in the experiment (.916 combined average). OKTF, on the other hand, exhibited very poor performance (.874 combined average). It was notably worse than many of measures in Table 4.

For parameters, we found that for OK, fixing $b = 1.0$, independent of k_1, was reasonable. Given this corroboration with the previous experiment, we argue that it makes sense to view OK as a single parameter function, having a fixed b and taking just a k_1 parameter. Again, similar to the previous experiment, we found that a wide range of k_1 values performed well. However, for this experiment OK was even less sensitive to k_1—beyond a threshold of approximately $k_1 = 8$, results were very similar. Fig. 5 gives an example of this behavior. The behavior of k_1 is consistent with previously the observed behavior of k_1 for OKC [20].

Our language modeling similarity measures performed well in the clustering experiment, with JMLM being 2nd in combined average (.917), and DLM being 1st (.922). Fig. 5 shows their combined average as their parameters vary. It is notable from the figure that JMLM's combined average is only mildly sensitive to λ. It varied by less than 1% over the entire range of λ we tested, although we observed a slight in peak at $\lambda = 0.7$. This insensitivity is consistent with that of the previous experiment. It makes using JMLM less worrisome in practice, as none-optimal λs seem to not degrade its results much. On the other hand, DLM's performance fluctuated much more noticeably, with its optimal range be-

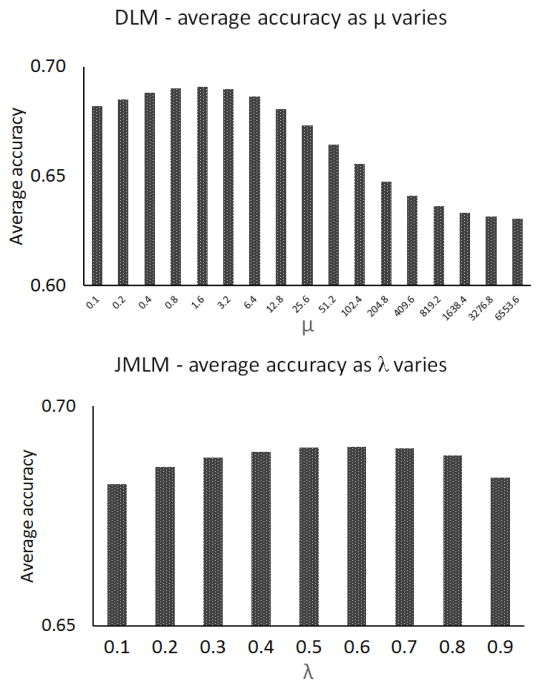

Figure 2: JMLM and DLM's average accuracy as their respective parameters change. μ is based on Manhattan length normalized documents.

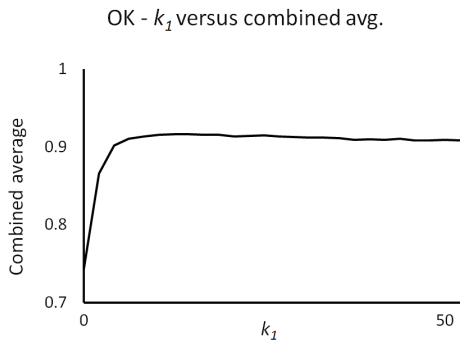

Figure 3: OK's combined average as k_1 varies and $b = 1.0$.

ing higher than the previous experiment (1.6 to 12.8 versus 0.8 to 3.2 previously). However, it also had a significantly higher combined average for all parameter settings in that range than the highest combined average from JMLM.

With respect to our DFR inter-document similarity measures, they performed slightly worse than in the previous experiment. Their best performer, DFR-In-L-EUC-COS (.907), was significantly worse than DLM, JMLM, and OK. However, three DFR measures were still significantly superior to TIC. We noticed that DFR measures using log length normalization performed poorly relative to the last experiment, whereas those using Euclidean length normalization ranked approximately the same as the previous experiment. For example, DFR-In-L-EUC-JAC was ranked 7th in the previous experiment, and 6th in this one.

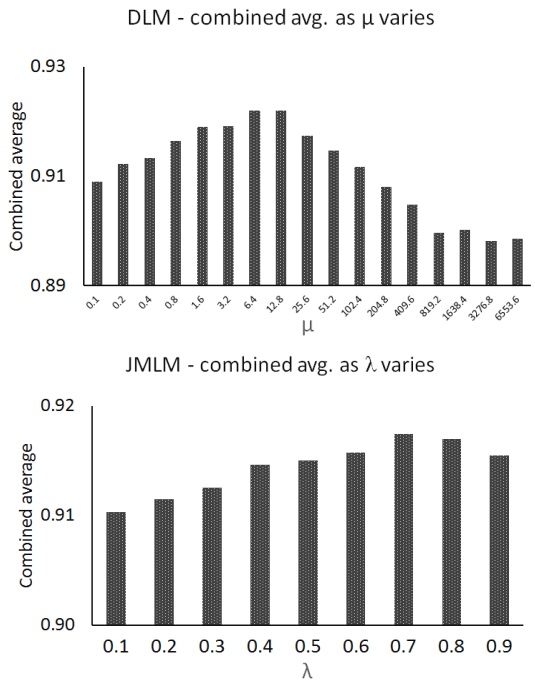

Figure 4: JMLM and DLM's combined averages as their respective parameters change. μ is based on Manhattan length normalized documents.

Considering just the methods in Table 3, OKC was, again, the best performer (.909). TIEN was the next best (.899), ranking 8th, with only a small (but still significant) difference between it and LIEN and TIC. That TIC was in the top 10 is noteworthy. Given that all the inter-document similarity measures that substantially outperformed TIC in our experiments were either designed in this paper, or only recently shown to be superior to it in clustering (OKC [20]), provides some rationale for its nearly ubiquitous use in clustering to date. All the other measures in Table 3 were notably worse than TIC (the next best was LIEN at .892).

Besides those results previously discussed, we found a few noteworthy points about the inter-document similarity measures in Table 3 from our experiments that we believe may apply in general:

1. Inter-document similarity measures using binary term weights yielded very poor results overall.

2. Euclidean and Manhattan length normalizations were never notably harmful to measures, and often substantially improved them. This provides some motivation for the widespread use of normalizations in inter-document similarity measures.

3. Including an idf component with a term frequency component was beneficial overall, but did not necessarily improve results for a specific dataset or clustering algorithm. The difference idf made was large.

4. The collection frequency component, as we implemented it, was universally inferior to the idf component.

5. Logging term frequency had only a small effect on inter-document similarity measures, except when they

used no length normalization. In this situation, it greatly improved results.

Summarizing the key result from the previous sections, our experiments indicate that OK, DLM, and JMLM are the best of all the inter-document similarity measures that we considered in this paper. Each has easily selected parameters and similar overall performance. They should replace cosine with tf-idf, as well as the other measures we considered, in clustering, as well as in general, as standard inter-document similarity measures. We also showed that best of our DFR measures were competitive with the best inter-document similarity measure we found in previous research, OKC (BM25 term-weighting with cosine). It should be noted that these conclusion are based on our specific datasets, each of which contains relatively long documents. It is possible that different similarity measures are better for shorter text sources such as tweets or forum posts. Additionally, length normalizations might need to be altered, or even omitted, in most of the similarity measures we considered in this paper to effectively handle the smaller length of such text sources. If this is the case is an avenue of future research.

6. CONCLUDING DISCUSSION

Despite query-document measures having moved beyond simple techniques such as cosine with tf-idf term weighting, the related area of inter-document similarity measure research still makes frequent use of such measures. In this paper, we focused on improving inter-document similarity measures by leveraging knowledge available from research on query-document measures. We implemented novel inter-document similarity measures based on BM25, language modeling, and divergence from randomness ranking functions, all of which are known to substantially and significantly outperform tf-idf with cosine in search.

We tested our novel inter-document similarity measures in a general experiment, as well as one focused on clustering specifically. The key result from those experiments was that OK, JMLM, and DLM are highly effective inter-document similarity measures, outperforming cosine with tf-idf term weighting, as well as a large variety of other inter-document similarity measures. Furthermore, our experiments show that reasonable universal parameter ranges exist for each of those measures, allowing them to be applied in practice more effectively. Together, these facts led us to conclude that they should replace the current standard inter-document similarity measures such as cosine with tf-idf term weighting.

One interesting avenue of future research is investigating previous research conclusions based on older inter-document similarity measures. For example, Zhao and Karypis [23, 24, 25] present a series of objective functions for document clustering which are evaluated using cosine with tf-idf term weighting. However, our preliminary tests suggest that when replicating their experiments with our better inter-document similarity measures, results will not only globally improve, but also change some conclusions drawn from their experiments. Specifically, the optimal objective functions change. We believe investigations like this one are necessary for inter-document similarity measures to improve.

Another possible avenue of research is the analysis of inter-document similarity measures in datasets where ground truth similarity is (or will be) defined (or tested) on a document pair level using concepts such as relevance. For example, in the TREC Web Track diversity task dataset[4], documents have relevance judgements with respect to subtopics. The overlap in subtopics might be viewed as a supervised inter-document similarity measure, and could be compared to unsupervised inter-document similarity measures using experiments similar to the ones we performed in this paper. Finally, we are considering the design of document clustering algorithms based on language modeling, as well as those exploiting learning-to-rank methods.

7. REFERENCES

[1] G. Amati and C. J. V. Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20:357–389, 2002.

[2] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 436–442, 2002.

[3] D. Boley, M. Gini, R. Gross, E.-H. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore. Document categorization and query generation on the World Wide Web using WebACE. *Artificial Intelligence Review*, 11:365–391, 1999.

[4] C. J. C. Burges, R. Ragno, and Q. V. Le. Learning to rank with nonsmooth cost functions. In *20th Annual Conference on Neural Information Processing Systems*.

[5] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender. Learning to rank using gradient descent. In *22nd International Conference on Machine Learning*, pages 89–96, 2005.

[6] C. L. A. Clarke, G. V. Cormack, and S. Büttcher. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010.

[7] B. C. M. Fung, K. Wang, and M. Ester. Hierarchical document clustering using frequent itemsets. In *Proceedings of the 3rd SIAM International Conference on Data Mining*, pages 59–70, 2003.

[8] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou. Exploiting Wikipedia as external knowledge for document clustering. In *Proceedings of the ACM SIGKDD 15th International Conference on Knowledge Discovery and Data Mining*, pages 389–396, 2009.

[9] R. Jin, C. Falusos, and A. G. Hauptmann. Meta-scoring: automatically evaluating term weighting schemes in ir without precision-recall. In *Proceedings of the ACM SIGIR 24th International Conference on Research and Development in Information Retrieval*, pages 83–89, 2001.

[10] T. Joachims. Optimizing search engines using clickthrough data. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[11] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

[12] J. W. Reed, Y. Jiao, T. E. Potok, B. A. Klump, M. T. Elmore, and A. R. Hurson. In *Machine Learning and Applications*, pages 258–263, 2007.

[4] `http://plg.uwaterloo.ca/~trecweb/2011.html`

[13] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *TREC '94: The Third Text REtrieval Conference*, 1994.

[14] X. Sevillano, G. Cobo, F. Alías, and J. C. Socoró. Feature diversity in cluster ensembles for robust document clustering. In *Proceedings of the 29th ACM SIGIR International Conference on Information Retrieval*, pages 697–698, 2006.

[15] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd ACM SIGIR International Conference on Informationa Retrieval*, pages 208–215, 2000.

[16] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Text Mining Workshop*, 2000.

[17] A. Strehl and J. Ghosh. Value-based customer grouping from large retail data-sets. In *SPIE Conference on Data Mining and Knowledge Discovery*, 2000.

[18] K. M. Svore and C. J. C. Burges. A machine learning approach for improved BM25 retrieval. In *18th ACM Conference on Information and Knowledge Management*, pages 1811–1814, 2009.

[19] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[20] J. S. Whissell and C. L. A. Clarke. Improving document clustering using Okapi BM25 feature weighting. *Information Retrieval*, 14:466–487, 2011.

[21] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th ACM SIGIR International Conference on Information Retrieval*, pages 267–273, 2003.

[22] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22:179–214, 2004.

[23] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. Technical Report 01-40, University of Minnesota, Department of Computer Science/Army HPC Research Center, 2001.

[24] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, pages 515–524, 2002.

[25] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55:311–331, 2004.

[26] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun. A general boosting method and its application to learning ranking functions for web search. In *21st Annual Conference on Neural Information Processing Systems*, pages 1697–1704, 2007.