Content-Based Relevance Estimation on the Web Using Inter-Document Similarities

Fiana Raiber
Technion, Israel
fiana@tx.technion.ac.il

Oren Kurland
Technion, Israel
kurland@ie.technion.ac.il

Moshe Tennenholtz
Microsoft Research and
Technion, Israel
moshet@ie.technion.ac.il

ABSTRACT

In adversarial and noisy search settings as the Web, the document-query surface level similarity can be a highly misleading relevance signal. Thus, devising *content-based* relevance estimation (ranking) approaches becomes highly challenging. We address this challenge using two methods that utilize inter-document similarities in an initially retrieved list. The first removes documents from the list that exhibit high query similarity, but for which there is insufficient additional support for relevance that is based on inter-document similarities. The method is based on a probabilistic model that decouples document-query similarities from relevance estimation. The second method re-ranks the list by "rewarding" documents that exhibit high similarity both to the query and to other documents in the list. Both methods incorporate, in addition, at the model level, query-independent document quality estimates. Extensive empirical evaluation demonstrates the merits of our methods.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

Keywords: Web search, inter-document similarities

1. INTRODUCTION

In many retrieval methods, the surface-level similarity of a document to a query is an important source of information in estimating the relevance of the document to the underlying information need. Over the Web, this "relevance indicator" can be highly misleading due to the adversarial and noisy nature of the retrieval setting; e.g., search engine optimization (SEO) efforts that employ content manipulation [6] can significantly bias relevance estimates that are based on surface level document-query similarities.

We present a novel probabilistic *content-based* relevance estimation model that accounts for the fact that document-query similarity can be an ineffective, and even (maliciously) misleading, indicator. The key idea is decoupling document-query similarities from relevance estimation; the latter is based on using inter-document-similarities in an initially retrieved list. The method that we devise, based on this model, removes documents from the initial list if they exhibit high surface-level query similarity that is not indicative of the ability to satisfy the information need — i.e., there is lack of sufficient additional content-based relevance evidence induced from inter-document similarities. A second method that we present re-ranks the initial list based on this relevance evidence. Both methods incorporate, at the model level, query-independent document quality estimates.

The Web-retrieval performance of our methods is substantially better than that of using only document-query similarities. The performance of our re-ranking method also transcends that of state-of-the-art methods that utilize (i) term proximity information [10], (ii) hyperlinks-based information, and (iii) content-based (query-independent) document quality measures [1]. In addition, we demonstrate the merits of integrating our methods with spam detection [4].

It is important to note that the content-based relevance estimates that we focus on should be viewed as one type of "feature" among the many used by Web search methods. The importance of this "feature" is that it directly touches on the definition of relevance; i.e., whether the document content satisfies the information need.

2. RELATED WORK

Our first method differs from classical retrieval paradigms by the virtue of decoupling document-query similarities from relevance estimation *at the model level*. This helps to account for the potential misleading nature of these similarities which might be due to adversarial effects. This is also an important difference between our model and previous work on using inter-document similarities. (See [5, 7] for surveys). Another difference between our two proposed methods and a previous suite of methods that utilize inter-document similarities [7] is treating inter-document similarities and query-independent document-quality estimates, *at the model level*, as complementary types of information rather than as substitutable [7]. Thus, our second (re-ranking) method essentially generalizes past approaches [7].

3. RETRIEVAL FRAMEWORK

Suppose that some search algorithm is employed over a corpus of documents $\mathcal{D}$ in response to query q to satisfy the presumed information need I that q expresses. As a result, an *initial ranking* of $\mathcal{D}$ is induced. We use $\mathcal{D}_q^{[k]}$ to denote the result list of the k most highly ranked documents.

We assume that the initial ranking is based on, among other criteria, the surface-level similarity between q and doc-

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$15.00.

uments. Naturally, then, the result list, $\mathcal{D}_q^{[k]}$, can contain documents not relevant to I that exhibit high surface-level similarity to q. In adversarial retrieval setting such as the Web, some of these documents might have been manipulated (e.g., keyword stuffed) to include substantial presence of q's terms. Others might simply bear very little textual content to begin with and/or provide only partial textual cover to I; e.g., a document containing a table that is composed only of a few keywords, some of which are q's terms [1].

3.1 A filtering approach

The goal of our first approach is filtering out from $\mathcal{D}_q^{[k]}$ documents that exhibit high surface-level similarity to q, but which do not pertain to the information need I.

Let $\mathcal{S}$ be the set of documents in the corpus that contain "suspiciously" high presence of q's terms; that is, presence that does not necessarily attest to the information needs that the document content can satisfy. (Refer to the keyword stuffing and table examples from above.) Let NR be the set of non-relevant documents in the corpus. Our goal is estimating for each document d in $\mathcal{D}_q^{[k]}$ the probability that it does not satisfy I (i.e., is not relevant), and has a "suspiciously" high presence of q's terms:

$$p(d \in \mathcal{S} \cap NR | q, I). \qquad (1)$$

More formally, let $\mathbf{q}$ and $\mathbf{I}$ denote random variables that take as values queries and information needs, respectively. Then, Equation 1 amounts to $p(d \in \mathcal{S}, d \in NR | \mathbf{q} = q, \mathbf{I} = I)$. Part (a) of Figure 1 presents the dependencies between random variables and events that we utilize. It is important to observe the following. Once $\mathbf{q}$ is set to the query q, $\mathcal{S}$ is uniquely determined; yet, $\mathcal{S}$ can correspond to various queries. Furthermore, by definition, the binary event $d \in \mathcal{S}$ is independent of $\mathbf{I}$, as it depends on the connection between d and q and properties of d itself. Another observation is that once we fix $\mathbf{I}$ to the specific information need I, the set of non-relevant documents, NR, is uniquely determined. Moreover, this set does not depend on $\mathbf{q}$ as relevance is determined only with respect to the information need. Then, one of our tasks is to estimate the probability that the binary event that represents non-relevance, $d \in NR$, happens. As is common, to facilitate notation we omit random variables from the formulation and use their values. Using the dependencies from part (a) of Figure 1 we get:

$$p(d \in \mathcal{S} \cap NR | q, I) = p(d \in \mathcal{S}|q) p(d \in NR|I). \qquad (2)$$

The separate treatment of (non-)relevance and surface-level document-query similarity *at the model level*, which is manifested in Equation 2, is a fundamental aspect by which our approach departs from classical retrieval paradigms. This separation helps target non-relevant documents that exhibit high surface-level query similarity. In an adversarial retrieval setting, some of these documents, for example, might have been stuffed with q's terms. Classical retrieval approaches, on the other hand, are often based, in spirit, on estimating $p(d \in R|I)$, the probability that d is in the set R of documents relevant to I. (Refer to part (b) of Figure 1 that we further discuss below.) Then, q is coupled with I, and $p(d \in R|q)$, which is often estimated based on the similarity between d and q, is used for ranking. The second aspect that differentiates our approach from classical retrieval methods is the actual estimates of (non-)relevance explored next.

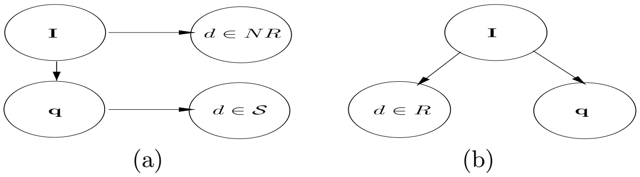

Figure 1: The dependencies used by the (a) Filter, and (b) ReRank methods.

3.1.1 Estimates

To derive a document scoring method from Equation 2, we first estimate $p(d \in \mathcal{S}|q)$. Since q is fixed, and consequently so is $\mathcal{S}$, we get

$$p(d \in \mathcal{S}|q) \overset{rank}{=} p(q|d \in \mathcal{S}) p(d \in \mathcal{S}). \qquad (3)$$

The probability $p(q|d \in \mathcal{S})$ can be interpreted as follows. Given document d, which is among those having a "suspiciously" high presence of terms from *some* query, what is the probability that the query at hand is q? (Recall that $\mathcal{S}$ can correspond to different queries.) A natural estimate is based on the similarity between q and d (we use two different similarity measures in Section 4.1):

$$\hat{p}(q|d \in \mathcal{S}) \propto sim(q, d). \qquad (4)$$

Here and after, $\hat{p}(x)$ denotes an estimate of $p(x)$.

The next step is estimating $p(d \in \mathcal{S})$ in Equation 3 . As q is not conditioned upon, and $\mathcal{S}$ can correspond to various queries, $p(d \in \mathcal{S})$ can be thought of as the prior probability of d having "suspiciously" high presence of terms — i.e., terms with presence not necessarily reflecting the information needs that d's content can satisfy. The probability that d is *spam* can serve, for example, for an estimate, as spam documents, by definition, do not satisfy any information needs. Other document-quality measures [7, 1] can serve to the same end and be integrated as we show in Section 4.2. One such measure, which we found to be quite effective in our experiments, is the *entropy* of the term distribution of d [7]. The term distribution is expected to have relatively high peaks, rather than be relatively flat, in case a document is "low on content", and has a non-representative occurrence of terms; in this case, the entropy is low. Formally, if $p_d(w)$ is the probability assigned to term w by a language model induced from d (details in Section 4.1), then d's entropy is defined as $H(d) \overset{def}{=} -\sum_{w \in d} p_d(w) \log p_d(w)$, and,

$$\hat{p}(d \in \mathcal{S}) \propto \exp(-H(d)). \qquad (5)$$

Estimating non-relevance. The second component of Equation 2 that we have to estimate is $p(d \in NR|I)$, the probability that d is not relevant to I. To that end, we leverage insights gained in work on re-ranking search results in "clean" settings (e.g., newswire collections) [5, 7]; specifically, that similarity to documents in a short list retrieved in response to a query can imply relevance [5, 7]. The reason, following the cluster hypothesis [12], is that similarity between relevant documents is potentially stronger than that between non-relevant documents, and that between relevant and non-relevant documents.

Accordingly, we assume that low similarity with documents in $\mathcal{D}_q^{[k]}$ indicates non-relevance. However, in contrast to work on utilizing inter-document-similarities in "clean" (e.g., newswire) settings [5, 7], which are free of content manipulation, and which often contain documents of high quality, here we compute inter-document similarities while disregarding query terms. The motivation is to account for the fact that query-terms occurrence is not necessarily indicative of the document content, a situation which is exacerbated in adversarial (and noisy) settings such as the Web. (Empirical exploration — actual numbers are omitted due to space considerations — supported the merits of disregarding query terms.) Formally, let $sim_{\bar{q}}(\cdot,\cdot)$ be an inter-document similarity measure disregarding q's terms (details in Section 4.1), then

$$\hat{p}(d \in NR|I) \propto \frac{1}{\sum_{d_i \in \mathcal{D}_q^{[k]} \setminus \{d\}} sim_{\bar{q}}(d, d_i)}. \qquad (6)$$

Using the estimates from above to instantiate Equation 2 yields our **Filter** method that scores d by[1]

$$S_{Filter}(d;q) \stackrel{def}{=} \frac{sim(q,d)\exp(-H(d))}{\sum_{d_i \in \mathcal{D}_q^{[k]} \setminus \{d\}} sim_{\bar{q}}(d, d_i)}.$$

The final corpus ranking is based on the initial ranking with the exception that the m documents d in $\mathcal{D}_q^{[k]}$ with the highest $S_{Filter}(d;q)$ are removed; m is a free parameter.

3.2 A re-ranking approach

An alternative approach that we consider is *re-ranking* $\mathcal{D}_q^{[k]}$ based on presumed relevance. Documents not in $\mathcal{D}_q^{[k]}$, i.e., those initially positioned at ranks lower than k, maintain their original ranks. Formally, for each d in $\mathcal{D}_q^{[k]}$ we estimate $p(d \in R|I)$, the probability that d is relevant to the information need I; R is the set of relevant documents in the corpus, which is uniquely determined once I is fixed. (Refer to part (b) of Figure 1.) To that end, we can use the rank equivalence $\hat{p}(d \in R|I) \stackrel{rank}{=} \hat{p}(I|d \in R)\hat{p}(d \in R)$.

We use the document term distribution entropy for the estimate of the prior of d's relevance, that is, $\hat{p}(d \in R)$. As noted above, high entropy might imply to a breadth of content. The set R can contain documents that also satisfy information needs other than I and which are represented by queries different than q; and, I itself can be represented by queries other than q. Thus, we use the following evidence combination for the estimate $\hat{p}(I|d \in R)$ of the probability that I is an information need satisfied by d. First, we measure the similarity between d and all other documents in $\mathcal{D}_q^{[k]}$. These documents can be considered "pseudo relevant" to I as they were retrieved in response to q which is a "signal" about I. Second, we scale the similarities just mentioned with the document-query surface level similarity [7]. This is intended to "balance" the effect that aspects not related to I have on the inter-document-similarities measured. Accordingly, our **ReRank** method scores d by:

$$S_{ReRank}(d;q) \stackrel{def}{=} sim(q,d)\exp(H(d)) \sum_{d_i \in \mathcal{D}_q^{[k]} \setminus \{d\}} sim_{\bar{q}}(d, d_i). \qquad (7)$$

4. EVALUATION

4.1 Experimental setup

In what follows we use language models for several purposes. Let $p_z^{Dir[\mu]}(\cdot)$ denote the Dirichlet-smoothed unigram language model induced from text z, with smoothing parameter μ [13]; unless otherwise stated, $\mu = 1000$ [13]. Setting $\mu = 0$ results in a non-smoothed maximum likelihood estimate that is used for (i) the corpus unigram language model, (ii) computing document entropy, and (iii) measuring inter-document similarities and language-model-based document-query similarity, as described below.

To measure inter-document similarities when disregarding q's terms, we use the cross entropy (CE) measure [8, 7]: $sim_{\bar{q}}(d_1, d_2) \stackrel{def}{=} \exp\left(-CE(p_{d_1}^{Dir[0]}(\cdot)||p_{d_2}^{Dir[\mu]}(\cdot))\right)$, with language models induced only *after* q's terms were removed from *all* documents in the corpus. The computational overhead incurred is not significant, as similarities are computed for at most a few hundreds top-retrieved documents.

To create an initial ranking (henceforth **init. rank.**) of the corpus in response to query q, we employ two methods. The first, denoted **LM**, is using $sim(q,d) \stackrel{def}{=} \exp\left(-CE(p_q^{Dir[0]}(\cdot)||p_d^{Dir[\mu]}(\cdot))\right)$, which amounts to the standard KL retrieval approach [8]; this document-query similarity measure is used in our methods when employed over the LM initial ranking. The second method, denoted **MRF**, is the Markov Random Field approach [10] employed with the sequential dependence model (SDM), which is a state-of-the-art content-based retrieval approach [10, 1]. Our methods use $sim(q,d) \stackrel{def}{=} p(d,q)$, which is induced using SDM, when operating on the MRF initial ranking. The free parameters of SDM, λ_T, λ_O and λ_U are set to 0.85, 0.1, and 0.05, respectively, as these values are known to yield effective performance [10, 1]. For the experiments with MRF, the document language model smoothing parameter, μ, was set to 2500 following previous recommendations [1].

For evaluation, we use the ClueWeb collection (category B) [3] that contains about 50 million documents. We use two query sets: 1-50 from TREC 2009 (**ClueWeb-09**), and, 51-100 from TREC 2010 (**ClueWeb-10**).

The Indri/Lemur toolkit (www.lemurproject.org) was used for experiments. Porter stemming was applied to both documents and queries. Stopwords appearing in a short list composed of 35 words [9] were removed only from queries. We use MAP (@1000), precision of the top 5 documents (p@5), NDCG (@20), and ERR (expected reciprocal rank) (@20) [2] for evaluation. Statistically significant performance differences are determined using the two-tailed paired t-test at a 95% confidence level.

Our Filter and ReRank methods operate on $\mathcal{D}_q^{[k]}$, the result list of k initially highest ranked documents. We set k to a value in $\{50, 100, 200, \ldots, 1000\}$; recall that Filter removes m documents from $\mathcal{D}_q^{[k]}$ and uses the initial ranking for the residual corpus; m is set to values in $\{1, 5, 10, 15, 20, 25\}$. The values of k and m, as those of the free parameters of the reference comparisons we use below, are set using 10-fold cross validation performed over queries; MAP serves for the optimization criterion in the training phase.

4.2 Experimental results

Main result. In Table 1 we present our main result for ClueWeb. Recall that the MRF initial ranking was cre-

[1] All the estimates in Filter can be turned to probability distributions by normalization with respect to documents. Note that such normalization does not affect ranking.

	ClueWeb-09								ClueWeb-10							
	LM				MRF				LM				MRF			
	MAP	p@5	NDCG	ERR	MAP	p@5	NDCG	ERR	MAP	p@5	NDCG	ERR	MAP	p@5	NDCG	ERR
init. rank.	18.0	27.2	25.2	27.0	19.5	26.8	26.5	25.7	14.6	13.3	12.3	8.0	14.3	13.8	12.1	6.3
OptMRF	20.4^i	31.2	30.3^i	31.9	20.6^i	30.8^i	30.4^i	31.1^i	15.3	13.3	13.1	7.8	15.4^i	13.3	13.1	7.8
Filter	18.8_o	39.2^i_o	29.9^i	35.4^i	20.8^i	40.4^i_o	32.7^i	37.6^i_o	15.4	21.2^i_o	16.0^i_o	11.7	16.7^i	26.2^i_o	18.2^i_o	11.9^i_o
ReRank	$\mathbf{21.6}^i$	$\mathbf{46.8}^i_o$	$\mathbf{33.8}^i$	$\mathbf{42.2}^i_o$	$\mathbf{22.5}^i$	$\mathbf{47.6}^i_o$	$\mathbf{34.7}^i$	$\mathbf{42.1}^i_o$	$\mathbf{18.6}^i_o$	$\mathbf{31.7}^i_o$	$\mathbf{23.9}^i_o$	$\mathbf{17.2}^i_o$	$\mathbf{20.2}^i_o$	$\mathbf{34.6}^i_o$	$\mathbf{26.8}^i_o$	$\mathbf{15.4}^i_o$

Table 1: Main result. Comparison with the initial ranking and OptMRF. Statistically significant differences with the former and the latter are marked with 'i' and 'o', respectively. Best result in a column is boldfaced.

ated by setting free parameters to some effective values. Hence, as a reference comparison we use MRF's sequential dependence model (SDM) for re-ranking the top 1000 documents retrieved by MRF and by LM, independently, where the free-parameter values are set using 10-fold cross validation; specifically, λ_T, λ_O and λ_U are set to values in $\{0, 0.05, 0.1, \ldots, 1\}$ where $\lambda_T + \lambda_O + \lambda_U = 1$. **OptMRF** denotes the resultant retrieval model as free-parameter values are optimized (with respect to MAP) in the training phase.

We see in Table 1 that none of the LM and MRF initial rankings posts performance that dominates that of the other across all relevant comparisons (query set × evaluation measure). Thus, we have initial rankings of different effectiveness with respect to the different evaluation measures. We also see that OptMRF outperforms in a vast majority of the relevant comparisons the LM and MRF initial rankings for both ClueWeb-09 and ClueWeb-10; for ClueWeb-09, many of the improvements are statistically significant.

We also see in Table 1 that Filter outperforms the initial ranking for all relevant comparisons; statistically significantly so in most cases. Our ReRank method is the best-performing in Table 1. It always outperforms the initial ranking, which is based only on document-query similarities, in a substantial and statistically significant manner; ReRank also outperforms OptMRF — often statistically significantly so and by quite a large margin; and, it outperforms Filter in all relevant comparisons. Further exploration (actual numbers are omitted due to space considerations) reveals that of the three information types used by ReRank, inter-document-similarities is the most effective one.

Applying spam detection. We now study the effectiveness of integrating our methods with spam detection. To that end, we first use as a reference comparison a common spam removal method denoted here **SpamRm** [4, 1]. Specifically, we apply Waterloo's spam detector [4] upon the initial ranking, top to bottom, and remove documents d with a non-spam score, $NonSpam(d)$, below 50 until we accumulate 1000 documents; $NonSpam(d)$ is in $[0, 100]$ and reflects the presumed percentage of documents in the entire ClueWeb English collection (category A, which includes around 500 million documents) that are "spammier" than d.

To integrate spam detection in Filter, we scale document d's score by $\frac{1}{NonSpam(d)+1}$, which amounts to using $\hat{p}(d \in \mathcal{S}) \propto \frac{\exp(-H(d))}{NonSpam(d)+1}$ instead of the original Equation 5; that is, the prior probability that d has a "suspiciously" high presence of terms that do not reflect the information needs that can be satisfied by d is now estimated based on *both* its term distribution entropy and the probability that d is spam. We use **Filter+SpRm** to denote the resultant retrieval model. Analogously, the score assigned to d by ReRank is scaled by $NonSpam(d) + 1$, which amounts to using both the document entropy and the probability that it is not spam as a prior for relevance. The resultant retrieval model is denoted **ReRank+SpRm**. Both Filter+SpRm and ReRank+SpRm operate on $\mathcal{D}_q^{[k]}$, as is the case for Filter and ReRank.

We see in Table 2 that applying suspected-spam removal (SpamRm) upon the initial ranking improves performance (except for MAP for ClueWeb-09). This finding is in line with previous reports [1]. We also see that ReRank improves over SpamRm in a majority of the relevant comparisons; for MAP, the improvements are often statistically significant. In the few cases ReRank is outperformed by SpamRm, the differences are not statistically significant. Filter, on the other hand, is outperformed by SpamRm in many cases.

We also see in Table 2 that integrating spam detection in our methods helps to somewhat improve their performance in many cases. (Compare "X+SpRm" with "X" rows.) Consequently, ReRank+SpRm is the most effective method in most relevant comparisons. Specifically, ReRank+SpRm outperforms SpamRm in a vast majority of the relevant comparisons, with quite a few of the improvements being statistically significant. These findings attest to the merits of integrating our methods with spam detection.

Document-quality measures. Filter and ReRank use the document entropy for a query-independent document quality measure [7, 1]. The fraction of stopwords in the corpus that appear on the page, and the ratio of stopwords to non-stopwords on the page, were shown to be even more effective as document-quality measures for Web retrieval [1]; and, the most effective quality measures among those explored [1]. (Here, a term is considered a stopword if it is among the 100 most frequent alphanumeric terms in the corpus [1].)

We use **ENT+SW** to denote a retrieval method that uses, in addition to query-document similarity, the entropy and the two stopwords-based features just described as document quality measures. In the language modeling case (LM), we use the document-quality values for scaling the document-query similarity, thereby treating the product of these values as a document relevance prior. In the MRF case, we follow recent work [1]. That is, we interpolate the quality values (after normalizing each to a $[0, 1]$ range) — having a different weight parameter assigned to each — with the value assigned to the document by the SDM model. The range of values for each such parameter is $\{0, 0.05, \ldots, 1\}$ with the constraint that the parameters' values, and those of the three SDM parameters mentioned above (λ_T, λ_O, and λ_U), sum to 1. Using the same approach, we study a method that integrates a document's **PageRank** score, which is a non-content-based quality measure, with the document-query similarity [1]. PageRank is computed upon the hyperlink graph of the English ClueWeb category A. (In the

	ClueWeb-09								ClueWeb-10							
	LM				MRF				LM				MRF			
	MAP	p@5	NDCG	ERR	MAP	p@5	NDCG	ERR	MAP	p@5	NDCG	ERR	MAP	p@5	NDCG	ERR
init. rank.	18.0	27.2	25.2	27.0	19.5	26.8	26.5	25.7	14.6	13.3	12.3	8.0	14.3	13.8	12.1	6.3
SpamRm	17.1	43.2^{i}	32.2^{i}	41.4^{i}	19.2	47.6^{i}	$\mathbf{36.8}^{i}$	$\mathbf{44.9}^{i}$	16.5	25.0^{i}	19.5^{i}	14.1^{i}	17.1^{i}	24.6^{i}	21.1^{i}	12.3^{i}
Filter	18.8	39.2^{i}	29.9^{i}	35.4^{i}	20.8^{i}	40.4^{i}	32.7^{i}	37.6^{i}_{s}	15.4	21.2^{i}	16.0^{i}_{s}	11.7	16.7^{i}	26.2^{i}	18.2^{i}	11.9^{i}
Filter+SpRm	19.8^{i}_{s}	42.8^{i}	32.4^{i}	38.6^{i}	22.2^{i}_{s}	46.8^{i}	35.7^{i}	43.9^{i}	15.5	20.4^{i}	16.1^{i}_{s}	13.5^{i}	16.1	22.9^{i}	17.5^{i}_{s}	12.3^{i}
ReRank	21.6^{i}_{s}	46.8^{i}	33.8^{i}	$\mathbf{42.2}^{i}$	$\mathbf{22.5}^{i}_{s}$	47.6^{i}	34.7^{i}	42.1^{i}	18.6^{i}	31.7^{i}	23.9^{i}	17.2^{i}	20.2^{i}_{s}	34.6^{i}_{s}	26.8^{i}	15.4^{i}
ReRank+SpRm	$\mathbf{21.8}^{i}_{s}$	$\mathbf{48.8}^{i}$	$\mathbf{35.5}^{i}$	42.1^{i}	22.4^{i}_{s}	$\mathbf{49.6}^{i}$	34.5^{i}	43.4^{i}	$\mathbf{18.7}^{i}$	$\mathbf{34.2}^{i}$	$\mathbf{26.0}^{i}$	$\mathbf{18.8}^{i}$	$\mathbf{20.8}^{i}_{s}$	$\mathbf{39.6}^{i}_{s}$	$\mathbf{29.3}^{i}_{s}$	$\mathbf{19.8}^{i}_{s}$

Table 2: Applying suspected-spam removal upon the initial ranking (SpamRm), and integrating spam detection in our methods (Filter+SpRm and ReRank+SpRm). Statistically significant differences with the initial ranking and SpamRm are marked with 'i' and 's', respectively; the best result in a column is boldfaced.

	ClueWeb-09								ClueWeb-10							
	LM				MRF				LM				MRF			
	MAP	p@5	NDCG	ERR	MAP	p@5	NDCG	ERR	MAP	p@5	NDCG	ERR	MAP	p@5	NDCG	ERR
init. rank.	18.0	27.2	25.2	27.0	19.5	26.8	26.5	25.7	14.6	13.3	12.3	8.0	14.3	13.8	12.1	6.3
PageRank	9.4^{i}	16.4^{i}	13.4^{i}	20.3	20.3	34.0^{i}	29.8^{i}	32.7^{i}	8.6^{i}	11.3	8.3^{i}	9.0	14.5	14.2	12.2	8.8
ENT+SW	16.3_{p}	30.0_{p}	25.1_{p}	26.7	21.7^{i}_{p}	37.6^{i}	32.0^{i}	31.9^{i}	15.7_{p}	18.3	22.8^{i}_{p}	10.9	16.5	15.8	19.6^{i}_{p}	9.5
ReRank	21.6^{i}_{pe}	46.8^{i}_{pe}	33.8^{i}_{pe}	42.2^{i}_{pe}	22.5^{i}	47.6^{i}_{pe}	34.7^{i}	42.1^{i}_{pe}	18.6^{i}_{pe}	31.7^{i}_{pe}	23.9^{i}_{p}	17.2^{i}_{pe}	$\mathbf{20.2}^{i}_{pe}$	$\mathbf{34.6}^{i}_{pe}$	$\mathbf{26.8}^{i}_{pe}$	15.4^{i}_{pe}
ReRank+SW	$\mathbf{22.1}^{i}_{pe}$	$\mathbf{49.6}^{i}_{pe}$	$\mathbf{36.3}^{i}_{pe}$	$\mathbf{46.3}^{i}_{pe}$	$\mathbf{22.9}^{i}_{p}$	$\mathbf{48.8}^{i}_{pe}$	$\mathbf{37.0}^{i}_{pe}$	$\mathbf{44.8}^{i}_{pe}$	$\mathbf{18.7}^{i}_{pe}$	$\mathbf{32.5}^{i}_{pe}$	$\mathbf{25.2}^{i}_{p}$	$\mathbf{17.5}^{i}_{pe}$	18.2^{i}_{p}	32.1^{i}_{pe}	24.6^{i}_{p}	$\mathbf{15.6}^{i}_{pe}$

Table 3: Comparison with using PageRank, and the entropy and stopwords-based indicators (ENT+SW), as document-quality measures; ReRank+SW is ReRank integrated with the stopwords-based measures; 'i', 'p' and 'e' mark statistically significant differences with the initial ranking, PageRank and ENT+SW, respectively. Boldface marks the best result in a column.

MRF case, the interpolation parameter for the PageRank score is set to a value in $\{0.05, 0.1, \ldots, 1\}$.) The ENT+SW and PageRank methods re-rank the 1000 initially highest ranked documents as recently proposed [1].

In Table 3 we compare the performance of PageRank, ENT+SW, ReRank, and, **ReRank+SW** — a variant of ReRank that uses both the entropy and the stopwords-based features as a relevance prior. (We use the product of the values; see Equation 7.) We can see that using PageRank as a document-quality measure helps improve performance over that of the initial ranking in the MRF case, but not in the LM case; these findings echo those from previous reports [11, 1]. In both cases, however, the performance of using PageRank is substantially worse than that of ReRank.

Table 3 shows that ReRank outperforms ENT+SW. Both methods utilize document-query similarity and document entropy. ENT+SW, in addition, uses the stopwords-based document quality values, while ReRank utilizes inter-document similarities. Thus, we see that inter-document similarities can be a more effective source of information than the stopwords-based document-quality measures, which are the most effective among those recently explored for Web search [1]. The fact that ReRank+SW improves over ReRank in many cases, except for MRF over ClueWeb-10, leads to the conclusion that query-independent document quality measures, and inter-document similarities, can be complementary sources of information for relevance estimation.

5. CONCLUSION

We presented two methods that address the challenge of content-based relevance estimation in adversarial (and noisy) retrieval setting as the Web, wherein document-query surface-level similarities can be a misleading relevance indicator. The methods utilize information induced from inter-document similarities in an initially retrieved list. Empirical evaluation demonstrated the merits of the methods.

Acknowledgments We thank the reviewers for their comments. This work has been supported by and carried out at the Technion-Microsoft Electronic Commerce Research Center. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

6. REFERENCES

[1] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proceedings of WSDM*, pages 95–104, 2011.

[2] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of CIKM*, pages 621–630, 2009.

[3] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *Proceedings of TREC*, 2009.

[4] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010.

[5] F. Diaz. Regularizing query-based retrieval scores. *Information Retrieval*, 10(6):531–562, 2007.

[6] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of AIRWeb*, pages 39–47, 2005.

[7] O. Kurland and L. Lee. Pagerank without hyperlinks: Structural reranking using links induced by language models. *ACM Transactions on Information Systems*, 28(4), 2010.

[8] J. D. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR*, pages 111–119, 2001.

[9] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[10] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proceedings of SIGIR*, pages 472–479, 2005.

[11] J. Peng, C. Macdonald, B. He, and I. Ounis. Combination of document priors in Web information retrieval. In *Proceedings of RIAO*, 2007.

[12] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, second edition, 1979.

[13] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR*, pages 334–342, 2001.