Re-ranking Documents Based on Query-Independent Document Specificity

Lei Zheng and Ingemar J. Cox

Department of Computer Science
University College London
London, WC1E 6BT, United Kingdom
lei.zheng@ucl.ac.uk, ingemar@cs.ucl.ac.uk

Abstract. The use of query-independent knowledge to improve the ranking of documents in information retrieval has proven very effective in the context of web search. This query-independent knowledge is derived from an analysis of the graph structure of hypertext links between documents. However, there are many cases where explicit hypertext links are absent or sparse, e.g. corporate Intranets. Previous work has sought to induce a graph link structure based on various measures of similarity between documents. After inducing these links, standard link analysis algorithms, e.g. PageRank, can then be applied. In this paper, we propose and examine an alternative approach to derive query-independent knowledge, which is not based on link analysis. Instead, we analyze each document independently and calculate a "specificity" score, based on (i) normalized inverse document frequency, and (ii) term entropies. Two re-ranking strategies, i.e. hard cutoff and soft cutoff, are then discussed to utilize our query-independent "specificity" scores. Experiments on standard TREC test sets show that our re-ranking algorithms produce gains in mean reciprocal rank of about 4%, and 4% to 6% gains in precision at 5 and 10, respectively, when using the collection of TREC disk 4 and queries from TREC 8 ad hoc topics. Empirical tests demonstrate that the entropy-based algorithm produces stable results across (i) retrieval models, (ii) query sets, and (iii) collections.

Keywords: Query-independent knowledge, Specificity, Normalized inverse document frequency, Entropy, Ranking, Information retrieval.

1 Introduction

It is now common for information retrieval to score documents based on a combination of query-dependent and query-independent information. Each resulting score is an estimate of the relevance of the document. The use of query-independent knowledge has proven particularly useful in the context of Web search [1,2,3,4]. Here, the graph structure created by the hypertext links between documents is used to estimate the "importance" of a document. Two well-known measures of document importance are Pagerank [1,2] and hyperlink-induced topic search (HITS) [3,4], which are discussed in detail in Section 2.

T. Andreasen et al. (Eds.): FQAS 2009, LNAI 5822, pp. 201–214, 2009.

These graph-based algorithms rely on links between documents. However, there are many collections, e.g. Intranets, where such links are absent or sparse. In these cases, it is often not possible to apply query-independent graph-based measures of document importance. To alleviate this problem, many researchers have proposed inducing a graph structure within the collection, based, for example, on the similarity between documents. This prior work is discussed in Section 2.

In this paper, we consider re-ranking documents based on documents' query-independent "specificity". Our fundamental assumption is that documents with a narrow focus (high specificity) are more important than documents with a broad focus (low specificity). We propose two measures of specificity based on (i) normalized inverse document frequency, and (ii) term entropies, as described in Section 3.

In Section 4, we describe a number of experiments using standard TREC test sets. The performance of the two specificity scores is compared. Subsequently, the stability of the entropy-based method is investigated with respect to different query sets and collections. Finally, we compare two different methods of combining query-dependent and query-independent scores. Section 5 then summarizes our results and discusses remaining issues.

2 Related Work

To paraphrase George Orwell [5], "All documents are equal but some documents are more equal than others". While several documents may have equal or similar query-dependent scores, significant improvements in retrieval are obtained by considering the query-independent "importance" of each document. Of course, the importance of a document can be quite subjective. And many factors may influence a document's importance. Considerable work has focused on approaches related to citation analysis. In particular, for Web documents, the links between documents are analogous to citations, and a number of graph-based link analysis algorithms have been proposed.

The most well-known measure of document importance is PageRank [1,2]. PageRank assigns every webpage a numerical score between 0 and 1, representing the likelihood that a person randomly clicking on links will arrive at a particular webpage. The score of PageRank is computed based on the link structure of the web graph. Berkhin [6], Langville and Meyer [2] investigated several methods for efficient computation of PageRank scores.

Kleinberg [3,4] proposed an alternative measure called hyperlink-induced topic search (HITS). The HITS algorithm assigns every webpage two scores. One is the hub score, and the other is the authority score. Generally, a webpage that *links to* many other webpages would be typically assigned a high hub score, and a webpage that *is linked to* by many other webpages would be typically assigned a high authority score. A systematic study of a number of HITS variants was conducted by Borodin *et al.* [7].

Both PageRank and HITS rely on links between documents. However, there are many collections where explicit hypertext links are absent or sparse. In these

cases, we can not directly apply link analysis algorithms. To overcome this limitation, Kurland and Lee [8,9] proposed inducing a graph structure for the top-k retrieved documents in response to a query. The k nodes of the induced graph are the top-k documents retrieved in response to a query. The weight on an edge between two nodes d_i and d_j is based on an estimation of the likelihood that if document d_i is relevant to the query, then d_j is also relevant. After constructing the graph, standard link analysis algorithms, e.g. PageRank and HITS, are then applied to re-rank the top-k retrieved documents. In [8], a method of structural re-ranking was discussed, and in [9], cluster-based language models are used for re-ranking. Specifically, in [9], Kurland and Lee reported a 4.6% gain for mean reciprocal rank (MRR), a 6.4% gain for precision at 5 (P@5), and a 4.8% gain for precision at 10 (P@10) based on the standard test set of TREC 8.

For other work on graph-based information retrieval, readers are directed to [10,11,12,13,14].

Our research differs from prior work in that our query-independent document score is not graph-based. Instead, we assume that documents with a narrow focus are more important than documents with a broad focus. We refer to the breadth of focus as "specificity". In the next Section, we propose two methods to estimate a document's specificity.

3 Document Specificity

We assume that documents containing unusual (specific) terms are more important than documents only containing common (broad) terms. To quantify this, we propose two specificity scores using statistical properties of the documents themselves. One is derived from the normalized inverse document frequency, and the other is based on the theory of information entropy.

3.1 Normalized IDF-Based Method

Inverse document frequency (IDF) is widely used as the measure of a term's discriminative ability. It is defined as the logarithmic ratio of total number of documents in a collection, n_d, to the number of documents containing the term (also known as term t_i's document frequency), $df(t_i)$, as shown in Equation 1 [15].

$$\mathrm{IDF}(t_i) = \log\left(\frac{n_d}{df(t_i)}\right) \tag{1}$$

We use normalized inverse document frequency (NIDF), as proposed by Robertson and Sparck-Jones [16]. The normalized IDF, defined in Equation 2, normalizes with respect to the number of documents not containing the term $(n_d - df(t_i))$ and adds a constant 0.5 to both the numerator and the denominator in order to moderate extreme values.

$$\mathrm{NIDF}(t_i) = \log\left(\frac{n_d - df(t_i) + 0.5}{df(t_i) + 0.5}\right) \tag{2}$$

Common words, such as "the", "and", "it", are likely to appear in every document within the collection and are therefore not discriminative. This poor discriminative capability is reflected in a correspondingly low NIDF value. Conversely, terms that only occur in a small number of documents are quite useful to discriminate between documents, and their NIDF values are correspondingly high.

Our assumption is that documents that consist primarily of terms with low NIDF values are less specific than documents that contain more discriminative terms. Under such assumption, we define a document specificity score, S_1, as:

$$S_1(d) = \frac{1}{l_d} \sum_{t_i \in d} tf(t_i) \text{NIDF}(t_i) = \frac{1}{l_d} \sum_{t_i \in d} tf(t_i) \log \left(\frac{n_d - df(t_i) + 0.5}{df(t_i) + 0.5} \right) \quad (3)$$

where $tf(t_i)$ is t_i's term frequency in document d, and l_d is the length of document d. The purpose of having a denominator l_d here is to reduce the influence of different document lengths.

3.2 Entropy-Based Method

The information entropy of a discrete random variable X with possible values $\{x_1, x_2, \cdots, x_n\}$ is defined as

$$H(X) = -\sum_{i=1}^{n} \Pr_X(x_i) \log \Pr_X(x_i) \quad (4)$$

where $\Pr_X(x_i)$ is the probability distribution function (pdf) of the random variable X.[1]

Entropy measures the uncertainty associated with the random variable X. Consider an example of a two-side coin. If the probability of the occurrence of either side is 1/2, the entropy achieves its maximum value, because we have the greatest uncertainty in the outcome (information content). However, if the probability for one side is 1/4 and for the other side 3/4, the uncertainty in the outcome reduces and the value of entropy reduces.

We consider each term t_i in the lexicon as a random variable. Term t_i is possibly occurring in document d_j, where j ranges from 1 to n_d. Therefore, the probability distribution of term t_i across the collection is

$$\Pr_{t_i}(d_j) = \frac{tf(d_j)}{tf(c)} \quad (j = 1, 2, \cdots, n_d) \quad (5)$$

where $tf(d_j)$ is t_i's term frequency in document d_j, and $tf(c)$ denotes t_i's term frequency in the whole collection c. Under such definition, the entropy of a term t_i is

$$H(t_i) = -\sum_{j=1}^{n_d} \Pr_{t_i}(d_j) \log \Pr_{t_i}(d_j) = -\sum_{j=1}^{n_d} \frac{tf(d_j)}{tf(c)} \log \left(\frac{tf(d_j)}{tf(c)} \right) \quad (6)$$

[1] In the case of $\Pr_X(x_i) = 0$ for some x_i, the value of the corresponding $0 \log 0$ should be taken to 0, which is given by the limit $\lim_{p \to 0} p \log p = 0$.

The probability, $\Pr_{t_i}(d_j)$, is the probability that a particular instance of the term, t_i, occurs in document d_j. If the term is a common word, e.g. "the", then the probability is almost the same for all documents (uniform distribution), and we have maximum uncertainty, i.e. a large entropy value. Conversely, if the term is unusual, e.g. "aardvark", then the probability is peaked, as only a few documents contain the term. In this case, the uncertainty is much less, and the entropy is correspondingly smaller. Note that the value of a term's entropy is inversely correlated to its normalized inverse document frequency. For NIDF, rare words have high values, whereas common words have low values.

After computing the entropy of each term t_i, our entropy-based measure of document specificity is given by

$$S_2(d) = \frac{1}{l_d} \sum_{t_i \in d} tf(t_i) H(t_i) = -\frac{1}{l_d} \sum_{t_i \in d} \left(tf(t_i) \sum_{j=1}^{n_d} \frac{tf(d_j)}{tf(c)} \log \left(\frac{tf(d_j)}{tf(c)} \right) \right) \quad (7)$$

Note that the higher the value of S_2, the less specific the document is. This is the inverse of our NDIF-based score.

4 Experimental Results

The use of both query-dependent and query-independent document scores requires the two scores to be combined to provide a final document score with which to rank documents. There are numerous methods to combine the two scores [17,8].

Here we considered a strategy in which each document was first classified as either "specific" or "unspecific" based on whether the document's specificity score was above or below a threshold. This classification was then used in one of two ways. In the first set of experiments, we simply remove all "unspecific" documents from our ranked list (*hard cutoff*). In the second set of experiments, the rank of "unspecific" documents is multiplied by an integer constant (*soft cutoff*). In both cases, performance is a function of the chosen threshold. Rather than reporting arbitrary threshold values, we report the percentage of documents in the collection that are classified as "unspecific", which is directly proportional to the threshold and provides a more meaningful value.

We use standard TREC collections in our experiments, as described in Table 1. These document collections do not contain link information. All our experiments are conducted using the LEMUR toolkit [18]. Documents are stemmed using the Krovetz stemmer [19]. We use the stopword list suggested by Fox [20], which

Table 1. Details of collections used in our experiments

Collection	Description	Number of documents
TREC disk 4	Federal Register (FR94) Financial Times (FT)	265,788
TREC disk 5	Federal Broadcast Information Service (FBIS) Los Angeles Times (LA)	262,367

includes a total of 421 stopwords. In all our experiments, the "title" part and the "description" part of TREC topics are used as evaluating queries.

4.1 Hard Cutoff

In this set of experiments, we consider the case where "unspecific" documents are removed from the ranked list. We refer to this as "hard cutoff", and define the hard cutoff rate as the percentage of documents in the collection that are classified as unspecific.

4.1.1 Comparison of NIDF-Based Method and Entropy-Based Method

We first compare our NIDF-based and entropy-based methods using the collection of TREC disk 4. TREC 8 ad hoc topics are used to evaluate the performance. Okapi BM25 [21] is used as the score function of the retrieval system.

Table 2. Comparison of NIDF-based method and entropy-based method at various hard cutoff rates based on MRR and its gain

Hard cutoff rate %	**0%**	**5%**	**10%**	**15%**	**20%**	**25%**
MRR (NIDF-based method)	0.6883	0.6984	0.6967	0.7005	0.7030	0.7033
Gain (NIDF-based method)	0	+1.47%	+1.22%	+1.77%	+2.14%	+2.18%
MRR (Entropy-based method)	0.6883	0.6984	0.6970	0.7004	0.7011	0.7108
Gain (Entropy-based method)	0	+1.47%	+1.26%	+1.76%	+1.86%	+3.27%
Hard cutoff rate %	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**
MRR (NIDF-based method)	0.7151	0.7042	0.7151	0.7059	0.6546	0.6669
Gain (NIDF-based method)	+3.89%	+2.31%	+3.89%	+2.56%	-4.90%	-3.11%
MRR (Entropy-based method)	0.7159	0.7058	0.6858	0.6885	0.6755	0.6620
Gain (Entropy-based method)	+4.01%	+2.54%	+0.36%	+0.03%	-1.86%	-3.82%
Hard cutoff rate %	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**
MRR (NIDF-based method)	0.6644	0.6363	0.5915	0.5803	0.5537	0.5012
Gain (NIDF-based method)	-3.47%	-7.55%	-14.06%	-15.69%	-19.56%	-27.18%
MRR (Entropy-based method)	0.6478	0.6387	0.6146	0.5971	0.5717	0.5167
Gain (Entropy-based method)	-5.88%	-7.21%	-10.71%	-13.25%	-16.94%	-24.93%

Table 2 is a comparison of mean reciprocal rank (MRR) and MRR gain. The reciprocal rank is the multiplicative inverse of the rank of the first relevant result. For example, if the first relevant result is ranked third in the response list (i.e. the first two documents are non-relevant), the reciprocal rank is 1/3. MRR is defined as the average of the reciprocal ranks to a set of queries, i.e.

$$\mathrm{MRR} = \frac{1}{n_q} \sum_{j=1}^{n_q} \frac{1}{\mathrm{Rank}_{1\mathrm{st,rel}}} \tag{8}$$

where n_q is the number of evaluating queries.

Table 2 suggests that both the NIDF-based method and the entropy-based method can provide improved performance compared with a ranking based only on a query-dependent score. For both methods, we obtain the best MRR gain

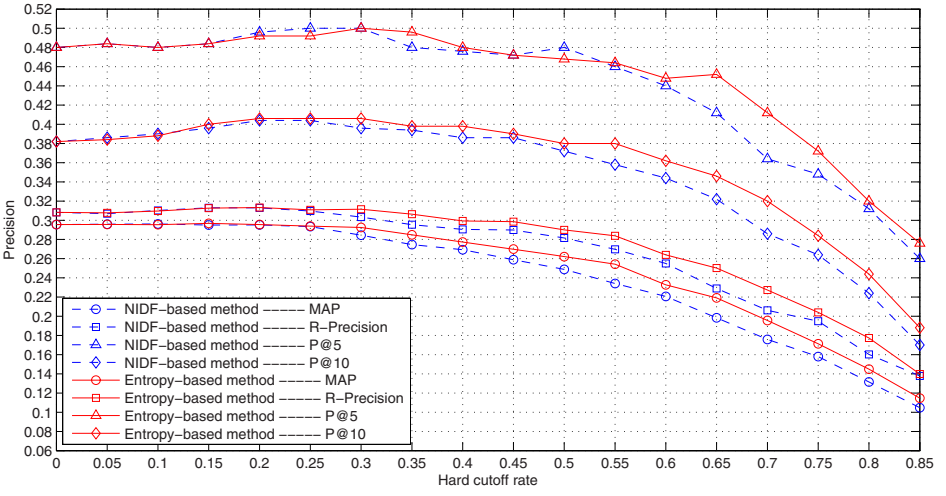

Fig. 1. Comparison of NIDF-based and entropy-based methods at various hard cutoff rates based on a variety of precision measures

(3.89% for the NIDF-based method and 4.01% for the entropy-based method) when the threshold is set such that the hard cutoff rate is 30%, i.e. 30% of the collection is classified as "unspecific".

Figure 1 compares our two methods based on various precision measures. Precision (P) is the fraction of retrieved documents that are relevant, i.e.

$$\mathrm{P} = \Pr(\text{relevant}|\text{retrieved}) = \frac{n_{\text{rel,ret}}}{n_{\text{ret}}} \tag{9}$$

Measuring precision at fixed levels of retrieved results, such as ten or thirty, is referred to as precision at k (P@k). Mathematically, it is the percent of retrieved documents that are relevant after k documents (whether relevant or not) have been retrieved, and then the values are averaged over all evaluating queries. P@k is important for many applications, since users may only examine the first page or the first few pages of the retrieved results. In this case, the quality of the top results becomes much more important. R-precision measures precision after R documents have been retrieved, where, for a given query, R is the total number of relevant documents in the collection. The average precision (AP) is the average of precisions after each relevant document is retrieved. The average of the AP values across all queries, is the mean average precision (MAP), i.e.

$$\mathrm{MAP} = \frac{1}{n_q}\sum_{j=1}^{n_q}\left(\frac{1}{n_{\text{rel}}}\sum_{k}\mathrm{P@Rank}_{k\text{th,rel}}\right) \tag{10}$$

where k is the rank of each relevant document to query q_j in the response list.

Figure 1 suggests that under almost all precision measures, the entropy-based method surpasses the NIDF-based method. There is nearly no precision

degradation when 20% to 30% of the collection is classified as "unspecific". Instead, for the entropy-based method, we obtain a 4.17% gain for P@5 (when 30% of the collection is classified as "unspecific"), a 6.28% gain for P@10 (30% "unspecific"), a 1.65% gain for R-Precision (20% "unspecific"), and a 0.41% gain for MAP (15% "unspecific").

Note that our experimental results are comparable to that of Kurland and Lee [9]. By using the method of inducing a graph structure, Kurland and Lee [9] obtained a 4.6% MRR gain, a 6.4% P@5 gain, and a 4.8% P@10 gain on their experiment of TREC 8 ad hoc topics. Note, however, that their document collection was a mixture of TREC disks 4 and 5.

Since the entropy-based method surpasses the NIDF-based method for almost all precision measures, we restrict further experiments to the entropy-based method only.

4.1.2 Performance Variation across Retrieval Models

Here we examine the sensitivity of our entropy-based specificity measure to different retrieval models. The document collection and the evaluating queries are the same as before. Three different retrieval models are examined. In addition to the Okapi BM25 probabilistic model [21], we considered another two widely used retrieval models, the Kullback-Leibler Divergence Language Model (LM) [22] and the classical term frequency-inverse document frequency (TFIDF) model [23].

Table 3 compares the three retrieval models based on MRR and MRR gain. For BM25 and TFIDF, we obtain the best MRR gain (4.01% and 5.02% respectively) when 30% of the collection is classified as "unspecific", while for the K-L

Table 3. Comparison of different retrieval models (Okapi BM25, K-L Divergence LM and Classical TFIDF) at various hard cutoff rates based on MRR and its gain

Hard cutoff rate %	**0%**	**5%**	**10%**	**15%**	**20%**	**25%**
MRR (Okapi BM25)	0.6883	0.6984	0.6970	0.7004	0.7011	0.7108
Gain (Okapi BM25)	0	+1.47%	+1.26%	+1.76%	+1.86%	+3.27%
MRR (K-L Divergence LM)	0.6166	0.6166	0.6179	0.6218	0.6231	0.6377
Gain (K-L Divergence LM)	0	0	+0.21%	+0.84%	+1.05%	+3.42%
MRR (Classical TFIDF)	0.6395	0.6310	0.6448	0.6537	0.6550	0.6621
Gain (Classical TFIDF)	0	-1.33%	+0.83%	+2.22%	+2.42%	+3.53%
Hard cutoff rate %	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**
MRR (Okapi BM25)	0.7159	0.7058	0.6858	0.6885	0.6755	0.6620
Gain (Okapi BM25)	+4.01%	+2.54%	+0.36%	+0.03%	-1.86%	-3.82%
MRR (K-L Divergence LM)	0.6366	0.6373	0.6305	0.6381	0.6524	0.6532
Gain (K-L Divergence LM)	+3.24%	+3.36%	+2.25%	+3.49%	+5.81%	+5.94%
MRR (Classical TFIDF)	0.6716	0.6481	0.6311	0.6245	0.6228	0.6263
Gain (Classical TFIDF)	+5.02%	+1.34%	-1.31%	-2.35%	-2.61%	-2.06%
Hard cutoff rate %	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**
MRR (Okapi BM25)	0.6478	0.6387	0.6146	0.5971	0.5717	0.5167
Gain (Okapi BM25)	-5.88%	-7.21%	-10.71%	-13.25%	-16.94%	-24.93%
MRR (K-L Divergence LM)	0.6315	0.6240	0.5714	0.5637	0.5239	0.4879
Gain (K-L Divergence LM)	+2.42%	+1.20%	-7.33%	-8.58%	-15.03%	-20.87%
MRR (Classical TFIDF)	0.6089	0.5978	0.5907	0.5940	0.5408	0.4976
Gain (Classical TFIDF)	-4.78%	-6.52%	-7.63%	-7.11%	-15.43%	-22.19%

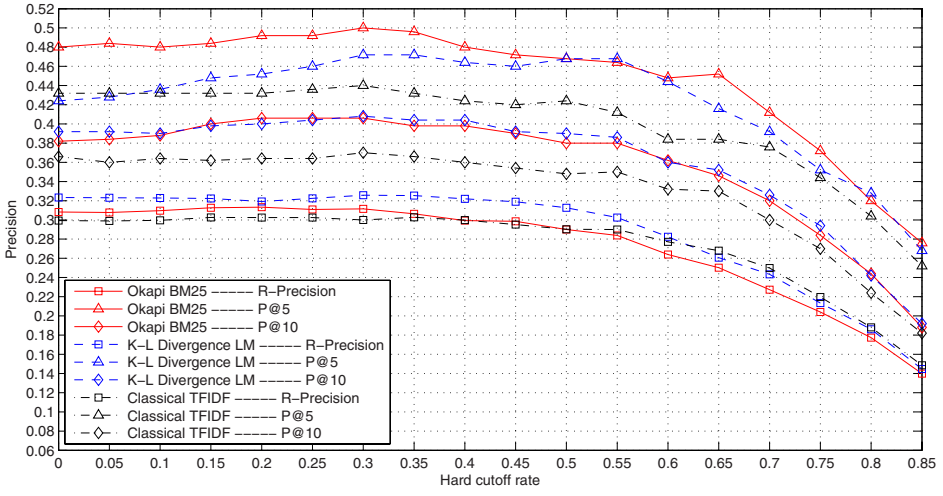

Fig. 2. Comparison of different retrieval models (Okapi BM25, K-L Divergence LM and Classical TFIDF) at various hard cutoff rates based on a variety of precision measures

Divergence LM, the best MMR gain (5.94%) occurs when 55% of the collection is classified as "unspecific". Note, however, that all three retrieval models exhibit improvements (4.01%, 3.24% and 5.02%) when 30% of the collection is classified as "unspecific".

Figure 2 shows various precision curves for the three retrieval models. Although the performance of the three retrieval models are different (due to the nature of the retrieval models themselves), our entropy-based method is stable across all three retrieval systems. Specifically, based on Okapi BM25, we obtain a 4.17% gain for P@5 and a 6.28% gain for P@10 when 30% of the collection is classified as "unspecific". Based on K-L Divergence LM, we obtain a 11.32% gain for P@5 and a 4.08% gain for P@10 when 30% of the collection is classified as "unspecific". Based on Classical TFIDF, we obtain a 1.85% gain for P@5 and a 1.09% gain for P@10 when 30% of the collection is classified as "unspecific". This provides some empirical evidence that performance improvements based on the specificity score are robust to various retrieval models.

4.1.3 Performance Variation across Query Sets

Here we examine the sensitivity of our entropy-based specificity measure across query sets. In previous experiments, we used ordinary queries (i.e. TREC 8 ad hoc topics). Here we test the performance on difficult queries[2]. The difficult queries are helpful for us to understand whether our measure of document specificity is stable for both the ordinary queries and difficult queries.

[2] In the TREC 2003 and 2004 robust tasks, NIST selected 50 difficult topics to evaluate the robustness (reliability) of a retrieval system.

Table 4. Comparison of ordinary queries and difficult queries on TREC disk 4 at various hard cutoff rates based on MRR and its gain

Hard cutoff rate %	**0%**	**5%**	**10%**	**15%**	**20%**	**25%**
MRR (Ordinary queries)	0.6883	0.6984	0.6970	0.7004	0.7011	0.7108
Gain (Ordinary queries)	0	+1.47%	+1.26%	+1.76%	+1.86%	+3.27%
MRR (Difficult queries)	0.5528	0.5648	0.5663	0.5734	0.5734	0.5763
Gain (Difficult queries)	0	+2.17%	+2.44%	+3.73%	+3.73%	+4.25%
Hard cutoff rate %	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**
MRR (Ordinary queries)	0.7159	0.7058	0.6858	0.6885	0.6755	0.6620
Gain (Ordinary queries)	+4.01%	+2.54%	+0.36%	+0.03%	-1.86%	-3.82%
MRR (Difficult queries)	0.5799	0.5752	0.5555	0.5507	0.5291	0.5541
Gain (Difficult queries)	+4.90%	+4.05%	+0.49%	-0.38%	-4.29%	+0.24%
Hard cutoff rate %	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**
MRR (Ordinary queries)	0.6478	0.6387	0.6146	0.5971	0.5717	0.5167
Gain (Ordinary queries)	-5.88%	-7.21%	-10.71%	-13.25%	-16.94%	-24.93%
MRR (Difficult queries)	0.5255	0.5088	0.4941	0.4802	0.4201	0.3923
Gain (Difficult queries)	-4.94%	-7.96%	-10.62%	-13.13%	-24.01%	-29.03%

Table 4 and Figure 3 summarize the experimental results on the collection of TREC disk 4 when using the Okapi BM25 [21]. For both ordinary and difficult queries, we obtain the best MRR gain when 30% of the collection is classified as "unspecific". A 4.01% MRR gain is obtained for ordinary queries, and a 4.90% MRR gain for difficult queries. For the precision curves shown in Figure 3, as expected, the absolute precision values of difficult queries decline significantly, since these are difficult queries. However, the relative performances generally remain similar. Specifically, based on ordinary queries, we obtain a 4.17% gain for P@5 and a 6.28% gain for P@10 when 30% of the collection is classified as "unspecific". Based on difficult queries, we obtain a 10.13% gain for P@5 when

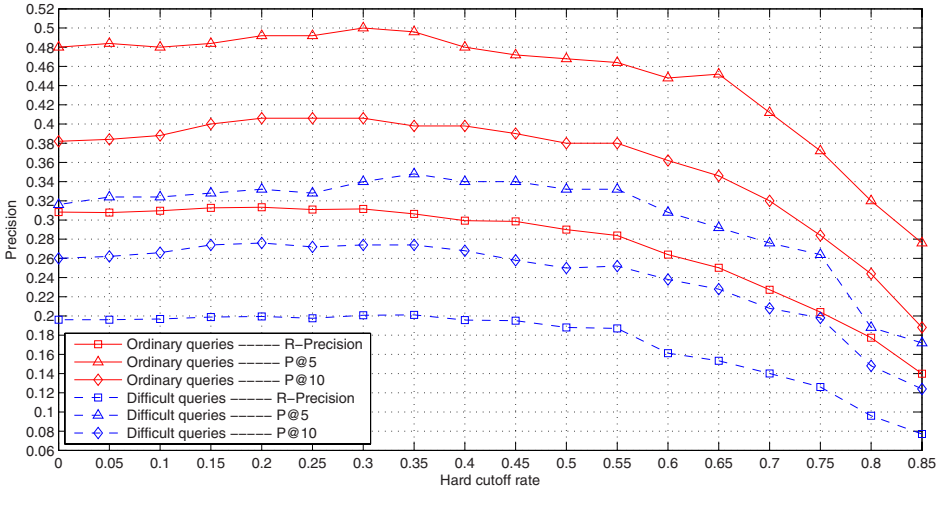

Fig. 3. Comparison of ordinary queries and difficult queries on TREC disk 4 at various hard cutoff rates based on a variety of precision measures

35% of the collection is classified as "unspecific", and a 6.15% gain for P@10 when 20% of the collection is classified as "unspecific".

4.1.4 Performance Variation across Collections

In addition to testing the sensitivity of our entropy-based specificity measure to retrieval models and query sets, we also examined the sensitivity across collections. In previous experiments, we used the TREC disk 4 as our document collection. Here we compare the experimental results on TREC disk 5.

Table 5. Comparison of ordinary queries and difficult queries on TREC disk 5 at various hard cutoff rates based on MRR and its gain

Hard cutoff rate %	**0%**	**5%**	**10%**	**15%**	**20%**	**25%**
MRR (Ordinary queries)	0.6218	0.6237	0.6301	0.6232	0.6295	0.6388
Gain (Ordinary queries)	0	+0.31%	+1.33%	+0.23%	+1.24%	+2.73%
MRR (Difficult queries)	0.4572	0.4602	0.4639	0.4563	0.4635	0.4846
Gain (Difficult queries)	0	+0.66%	+1.47%	-0.20%	+1.38%	+5.99%
Hard cutoff rate %	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**
MRR (Ordinary queries)	0.6380	0.6570	0.6435	0.6588	0.6489	0.6445
Gain (Ordinary queries)	+2.61%	+5.66%	+3.49%	+5.95%	+4.36%	+3.65%
MRR (Difficult queries)	0.4686	0.4841	0.5014	0.5071	0.4972	0.5018
Gain (Difficult queries)	+2.49%	+5.88%	+9.67%	+10.91%	+8.75%	+9.76%
Hard cutoff rate %	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**
MRR (Ordinary queries)	0.6223	0.6107	0.5947	0.5459	0.4583	0.4272
Gain (Ordinary queries)	+0.08%	-1.79%	-4.36%	-12.21%	-26.29%	-31.30%
MRR (Difficult queries)	0.5069	0.4964	0.4398	0.4023	0.3329	0.3111
Gain (Difficult queries)	+10.87%	+8.57%	-3.81%	-12.01%	-27.19%	-31.96%

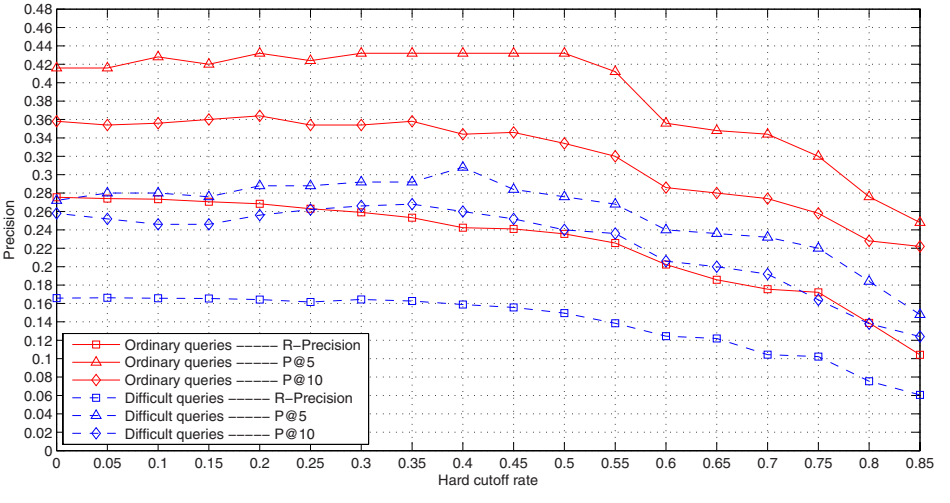

Fig. 4. Comparison of ordinary queries and difficult queries on TREC disk 5 at various hard cutoff rates based on a variety of precision measures

Table 5 and Figure 4 summarize the experimental results on the collection of TREC disk 5[3] when using the Okapi BM25 [21]. For both ordinary and difficult queries, we obtain the best MRR gain when 45% of the collection is classified as "unspecific". A 5.95% MRR gain is obtained for ordinary queries, and a 10.91% MRR gain for difficult queries. For the precision curves shown in Figure 4, the relative performances are similar to our previous experiment using the TREC disk 4. Specifically, based on ordinary queries, we obtain a 3.85% gain for P@5 and a 1.68% gain for P@10 when 20% of the collection is classified as "unspecific". Based on difficult queries, we obtain a 13.24% gain for P@5 when 40% of the collection is classified as "unspecific", and a 3.88% gain for P@10 when 35% of the collection is classified as "unspecific".

4.2 Soft Cutoff

The experiments of Section 4.1 used a "hard cutoff" strategy, in which documents classified as "unspecific" were removed from the ranked list. Note that "unspecific" documents cannot be retrieved under the hard cutoff strategy. In order to overcome this limitation, we consider an alternative strategy in which the query-dependent document ranks are weighted by some function of the query-independent document scores. Here, we report performance on such a "soft cutoff" strategy.

Our soft cutoff strategy multiplies the query-dependent ranks of "unspecific" documents by a factor of two. For example, if an "unspecific" document is initially ranked 4th, its rank is increased to $4 \times 2 = 8$. In the case where the final rank is greater than the length of the ranked list, the rank of the "unspecific" document will be increased to the bottom of the ranked list. Okapi BM25 [21] is once again used as the query-dependent score function of the retrieval system.

Page limitations prohibit enumerating comprehensive experimental results similar to those of Section 4.1. Instead, we report results for a threshold setting

Table 6. Comparison of original ranking, hard cutoff strategy and soft cutoff strategy. For each precision measure, the best result is given in italic.

Collection	Query	P@5			P@10		
		Original	**Hard cutoff**	**Soft cutoff**	**Original**	**Hard cutoff**	**Soft cutoff**
Disk 4	Ordinary	0.4800	*0.5000*	0.4880	0.3820	*0.4060*	0.3920
	Difficult	0.3160	*0.3400*	0.3240	0.2600	*0.2740*	0.2660
Disk 5	Ordinary	0.4160	*0.4320*	0.4240	0.3580	0.3540	*0.3660*
	Difficult	0.2720	*0.2920*	0.2800	0.2580	*0.2660*	0.2640
Collection	**Query**	**R-precision**			**MAP**		
		Original	**Hard cutoff**	**Soft cutoff**	**Original**	**Hard cutoff**	**Soft cutoff**
Disk 4	Ordinary	0.3082	0.3115	*0.3127*	0.2956	0.2925	*0.2978*
	Difficult	0.1961	*0.2008*	0.1985	0.1567	*0.1614*	0.1607
Disk 5	Ordinary	0.2755	0.2590	*0.2784*	0.2588	0.2322	*0.2589*
	Difficult	0.1658	0.1644	*0.1719*	0.1406	0.1298	*0.1417*

[3] Here we only report the results on the TREC disk 5, since the results on the TREC disk 4 was reported in Section 4.1.3.

where 30% of the collection are classified as "unspecific". Table 6 summarizes the experimental results.

Table 6 suggests that the hard cutoff strategy is superior when performance is based on the precision of the top-k retrieved documents, e.g. P@5 and P@10. However, for R-precision and MAP, the soft cutoff is generally superior. This may be because overall retrieval performance, as measured by R-precision and MAP, is likely to be more affected by the fact that the hard cutoff strategy discards all "unspecific" documents, than constrained retrieval performance, e.g. P@k.

5 Conclusions and Future Work

The use of query-independent knowledge to re-rank retrieved documents has previously been studied based on an explicit or implicit analysis of the graph structure between documents. In this paper, an alternative approach to derive query-independent knowledge is investigated, which is not based on link analysis. We assume that documents with a narrow focus are generally more relevant than documents with a broad focus, and propose two measures of this document "specificity". The two measures are based on normalized inverse document frequency and term entropies, respectively.

In our first set of experiments, documents were classified as either "specific" or "unspecific", and the latter were removed from the retrieval list. We referred to this as "hard cutoff". Experiments on the collection of TREC disk 4 and queries drawn from TREC 8 ad hoc topics showed that our re-ranking algorithms produce gains in mean reciprocal rank of about 4%, and 4% to 6% gains in precision at 5 and 10, respectively. The entropy-based specificity measure performed slightly better than that based on NIDF. Subsequent empirical tests with the entropy-based method produced stable results across (i) retrieval models, (ii) query sets, and (iii) collections. Further experimentation is recommended to verify this over a more varied set of parameters.

The hard cutoff strategy is equivalent to discarding "unspecific" documents from the collection. As such, "unspecific" documents can never be retrieved. To address this limitation, we also considered a "soft cutoff" strategy, in which documents classified as "unspecific" were not removed from the retrieval list, but had their rank increased. Experimental results showed that our soft cutoff strategy is superior on the overall retrieval performance, e.g. MAP. However, the precision gains based on top retrieved documents, e.g. P@5, favor the hard cutoff strategy.

In future work, we plan to investigate more sophisticated soft cutoff strategies, based on a Bayesian formulation. We will also try to refine our measures of document specificity and provide a thorough comparison with graph-based approaches.

Acknowledgments

The authors acknowledge valuable discussions with Jun Wang and Jianhan Zhu of University College London, especially Jun Wang's suggestion of "document specificity" to describe our query-independent NIDF and entropy-based document scores.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the 7th WWW, pp. 107–117 (1998)
2. Langville, A.N., Meyer, C.D.: Google's pagerank and beyond: the science of search engine rankings. Princeton University Press, Princeton (2006)
3. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: Proceedings of the 9th Symposium on Discrete Algorithms, pp. 668–677 (1998)
4. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604–632 (1999)
5. Orwell, G.: Animal Farm: A Fairy Story. Secker and Warburg (1945)
6. Berkhin, P.: A survey on pagerank computing. Internet Mathematics 2(1), 73–120 (2005)
7. Borodin, A., Roberts, G.O., et al.: Finding authorities and hubs from link structures on the world wide web. In: Proceedings of the 10th WWW, pp. 415–429 (2001)
8. Kurland, O., Lee, L.: Pagerank without hyperlinks: structural re-ranking using links induced by language models. In: Proc. of the 28th SIGIR, pp. 306–313 (2005)
9. Kurland, O., Lee, L.: Respect my authority!: Hits without hyperlinks, utilizing cluster-based language models. In: Proceedings of the 29th SIGIR, pp. 83–90 (2006)
10. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the 7th SIGKDD, pp. 269–274 (2001)
11. Joachims, T.: Transductive learning via spectral graph partitioning. In: Proceedings of ICML, pp. 290–297 (2003)
12. Liu, X., Croft, W.B.: Cluster-based retrieval using language models. In: Proceedings of the 27th SIGIR, pp. 186–193 (2004)
13. Zhang, B., Hua, L., et al.: Improving web search results using affinity graph. In: Proceedings of the 28th SIGIR, pp. 504–511 (2005)
14. Balinski, J., Danilowicz, C.: Re-ranking method based on inter-document distances. Information Processing and Management 41(4), 759–775 (2005)
15. Sparck-Jones, K.: Index term weighting. Information Storage and Retrieval 9(11), 619–633 (1973)
16. Robertson, S.E., Sparck-Jones, K.: Relevance weighting of search terms. Journal of the American Society for Information Science 27(3), 129–146 (1976)
17. Diaz, F.: Regularizing ad hoc retrieval scores. In: Proceedings of the 14th CIKM, pp. 672–679 (2005)
18. Ogilvie, P., Callan, J.: Experiments using the lemur toolkit. In: Proceedings of TREC-10 (2001)
19. Krovetz, R.: Viewing morphology as an inference process. In: Proceedings of the 28th SIGIR, pp. 191–202 (1993)
20. Fox, C.: A stop list for general text. SIGIR Forum 24(1-2), 19–21 (1990)
21. Sparck-Jones, K., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval. Information Processing and Management 36(6), 779–808 (2000)
22. Zhai, C.: Notes on the kl-divergence retrieval formula and dirichlet prior smoothing (2007)
23. Salton, G.: The SMART Retrieval System. Prentice-Hall, Inc., Englewood Cliffs (1971)