UMass at TREC 2010 Web Track: Term Dependence, Spam Filtering and Quality Bias

Michael Bendersky, David Fisher, W. Bruce Croft
Center for Intelligent Information Retrieval,
University of Massachusetts Amherst

Abstract

Many existing retrieval approaches treat all the documents in the collection equally, and do not take into account the content quality of the retrieved documents. In our submissions for TREC 2010 Web Track, we utilize quality-biased ranking methods that are aimed to promote documents that potentially contain high-quality content, and penalize spam and low-quality documents. Our experiments with the ad hoc web topics from TREC 2010 show that features such as the spamminess of the document (as computed by the Waterloo team [6]) and the readability of the document (modeled by the fraction of stopwords in the document) are very important for improving the precision at the top ranks. Promotion of the high-quality Wikipedia pages leads to further retrieval performance improvements. In addition, we found that using Wikipedia as a high-quality document collection for query expansion can ameliorate some of the negative effects of performing pseudo-relevance feedback from a noisy web collection such as ClueWeb09.

1 Introduction

In this notebook, we describe the details of the participation of the UMass Amherst team in the *adhoc task* of the TREC 2010 Web Track. Our participating systems focused on three major features of information retrieval on web scale: (a) *query term dependence*, (b) *spam filtering*, and (c) *document quality bias*. The first two features (term dependence and spam filtering) have been previously recognized as an important part of web search systems by past TREC participants, while the third one (document quality bias) is not as well explored and understood. Next, we describe these three features in more depth.

1.1 Term Dependence

Several teams at Web Track 2009 (for instance, [15, 7]) used the Markov Random Field retrieval model (MRF-IR) [13] to model term dependencies. Similarly, several teams at the Terabyte Track [5] and the Million Query Track [1] that preceded the Web Track used the MRF-IR model to build highly effective retrieval systems.

1.2 Spam Filtering

Spam detection and filtering was recognized as an important component of web retrieval in Web Track 2009. Lin et al. [12] used a commercial spam detection system to improve the quality of their submissions. Cormack et al. [6] created a publicly available *Waterloo Spam Ranking for the ClueWeb09 Dataset*, which was found to drastically improve the performance of practically all TREC Web Track 2009 submissions, when applied as a spam filter.

1.3 Quality Bias

In addition to the term dependence and spam detection features, which were incorporated in our submitted retrieval systems, we also tackled the issue of document quality bias that was recently found to improve the performance of web retrieval, even after the application of a spam filter and a link-based prior [3].

A quality of a web page is determined by a combination of many distinct factors. First, it has to contain original, trustworthy, and up-to-date content of genuine value. It should also provide metadata that accurately describes the content of a page, and contain links that can point people to other related resources. Finally, web page layout should be consistent and follow the principles of user-centric web design, by allowing readers to effortlessly navigate to the relevant information on the page. As document quality is influenced to some degree by all of these factors, the quality of a page should not be viewed as a dichotomy, but rather as a continuous spectrum.

At one end of this quality spectrum are well known resources for high-quality web documents such as Wikipedia. Wikipedia articles are constantly monitored and updated by editors, have a consistent layout and usually contain links to other related Wikipedia articles and web pages of interest. On the other end of this spectrum are *spam pages* that employ techniques such as content duplication, link schemes, content cloaking and keyword stuffing to artificially inflate their search engine ranking and provide no useful content (or even fraudulent and harmful content) to their readers.

Most of the pages on the web, however, are somewhere in between these two extremes on the quality spectrum. Many web pages do not have the same level of editorial supervision as Wikipedia, and might contain some outdated information, but still provide useful content to their readers. Many of the web pages also do not have a consistent easy-to-follow layout, making it harder to locate relevant portions of the text. However, these pages of lesser quality are still relevant to some user queries. This is especially true for rare and "niche" user information needs that often lack proper coverage by high quality resources such as Wikipedia. Therefore, we believe that it is important to explicitly incorporate the information about the quality of the page into the ranking produced by the search engine.

2 Modeling Term Dependence

In order to model the term dependence, we make use of the *sequential dependence* variant of the MRF-IR model proposed by Metzler and Croft [13]. Next, we briefly describe the basics of this model.

The sequential dependence model goes beyond single query terms, and incorporates both exact phrase matches (the exact phrase appears in the document) and proximity matches (the terms in the phrase appear in close proximity to each other in the document) into the retrieval function. The sequential dependence model considers the adjacent bigrams in the query. For instance, for a query `income tax return online` the following bigrams will be considered: `income tax`, `tax return`, and `return online`. The term and bigram exact and proximity matches are assigned different weights, which are tuned to optimize the retrieval performance. Using the Indri query language [16], the example query will be rewritten as

```
#weight( 0.8 #combine(income tax return online)
         0.15 #combine( #1(income tax) #1(tax return) #1(return online) )
         0.05 #combine( #uw8(income tax) #uw8(tax return) #uw8(return online) )
       )
```

where the weights for the different query parts are set based on the recommendations by Metzler and Croft [13]. We refer to the score assigned to the document by the sequential dependence model as $sc_{TD}(D)$.

3 Spam Filtering Stage

We used the spam filtering as described by Cormack et al. [6]. Using the publicly available *Waterloo Spam Ranking for the ClueWeb09 Dataset*. This ranking assigns a percentile $p(D)$ to each of the $\sim$ 500 million documents in the collection. We filter out the bottom 60% of the documents, as determined by the spam ranking, as this number was found to optimize ERR@10 and NDCG@10 in our preliminary experiments with the Web Track 2009 queries. That is the documents are scored by

$$sc_{SF}(D) = \begin{cases} sc_{TD}(D) & \text{if } p(D) \geq 60 \\ -\infty & \text{else.} \end{cases}$$

4 Quality Biasing Stage

While there is an abundance of features that can be associated with document quality (see, for instance, [3, 8, 14, 17] for an extensive examination of such features) for our submission we chose to focus on two features that we found to have a highly positive effect on the retrieval performance in our preliminary experiments with the Web Track 2009 queries.

- $I_w(D)$ — Wikipedia Indicator. Indicates whether the document D is a part of the `en.wikipedia.org` domain. Wikipedia pages are known to provide good answers to many web search queries, and are often promoted to the top ranks by the commercial search engines. An examination of the relevance judgments from Web Track 2009 reveals that more than 12% of the relevant documents are Wikipedia pages, which is much higher than the proportion of the Wikipedia pages in ClueWeb09. For instance, University of Amsterdam TREC team in Web Track 2009 achieved an impressive retrieval performance by simply promoting Wikipedia pages to the top ranks [7].
- $\sigma(D)$ — Stopword to non-stopword ratio of the document D. Percentage of the terms on the page that are in the *stopword list*. The stopword list is constructed using the top-100 most frequent alphabetic unigrams in a large web corpus [4]. Low values of $\sigma(D)$ usually characterize documents that have very poor readability and are very unlikely to contain useful and relevant information on any topic [9, 14].

The *quality biased* document score is computed, based on the two features above, as

$$sc_{QB}(D) = \begin{cases} \lambda sc_{SF}(D) + (1-\lambda) I_w & \text{if } \sigma(D) \geq \Phi \\ -\infty & \text{else,} \end{cases} \quad (1)$$

where λ and Φ are free parameters that are set to 0.8 and 0.1, respectively, since these values resulted in the highest ERR@10 and NDCG@10 in our preliminary experiments with the Web Track 2009 queries.

5 Query Expansion with High Quality Documents

In addition to biasing the retrieval towards high-quality documents by explicit promotion of the Wikipedia pages, we also employed a pseudo-relevance feedback technique, which used the Wikipedia pages to improve the effect of the pseudo-relevance feedback on ClueWeb09. We found that our initial pseudo-relevance runs on the Web Track 2009 queries failed to improve performance, due to the noisy nature of the retrieved set.

For this purpose, we separately indexed the Wikipedia pages, and used this index to expand the sequential dependence model queries with 20 additional terms. The expansion terms were selected using the standard Relevance Model [11]. Next, the expanded queries were run over the entire ClueWeb09 index with the spam filter and the quality bias (see Equation 1) applied.

6 Analysis of the Runs

6.1 Setup

The English portion of the ClueWeb09 corpus (a.k.a. Category A), was indexed using the Indri toolkit[1]. Both the index and the queries were stopped using a standard INQUERY stopwords list [2] and stemmed using a Krovetz stemmer [10].

All the retrieval experiments were performed using Indri as well. Indri query language was used to create the sequential dependence model queries, as well as to perform spam filtering and query expansion. The quality biasing scoring $sc_{QB}(D)$, described in Section 4, was applied to the top-10K results retrieved by the spam-filtered scoring $sc_{SF}(D)$.

6.2 Comparison of the Runs

In this section we compare between the three runs submitted to the adhoc task of the TREC 2010 Web Track. The runs we submitted were variations of the quality-biased ranking method described in the previous sections. The description of the three runs is as follows:

1. *SD* — This run scored the documents based on Equation 1, with λ set to 1. The purpose of this run was to evaluate the effect of the filtering stages (spam filtering and filtering of low-quality pages with very few stopwords) on the retrieval performance. (Submitted run ID — *umassSDM*.)
2. *SD+FB* — This run employed the pseudo relevance feedback based query expansion, as described in 5. The purpose of this run was to evaluate the effect of query expansion with high quality Wikipedia documents on the retrieval performance. (Submitted run ID — *umasswfb520*.)
3. *SD+W* — This run scored the documents using the Equation 1, with λ set to 0.8. The purpose of this run was to examine the effect of the promotion of the high quality Wikipedia pages on the retrieval performance. (Submitted run ID — *umassSDMW*.)

Table 6.2 compares the retrieval performance of these runs. It is clear that both expansion from Wikipedia and promotion of Wikipedia pages have a significant positive impact on the retrieval performance, even after filtering out spam and low-quality pages. Overall, the promotion strategy of the *SD+W* run seems to be preferable to the expansion strategy of the *SD+FB* run, especially given the fact that it does not require a computationally expensive procedure of query expansion.

[1] Available at `http://sourceforge.net/projects/lemur/`

	Binary Metrics		
	prec@5	b-pref	MAP
SD	32.50	22.80	11.35
SD+FB	**48.75***	**25.41**	14.04
SD+W	47.50*	24.29	**14.82***

	Graded Metrics	
	NDCG@10	ERR@10
SD	15.85	10.08
SD+FB	25.53*	12.72
SD+W	**26.87***	**12.80***

Table 1: **Comparison of the retrieval performance of the three submitted runs. Statistically significant differences (Wilcoxon sign test, $\alpha < 0.05$) with *SD* are marked by *. Best result per column appears in boldface.**

	ERR@10	NDCG@10
$\geq$ median	73%	71%
= min	23%	23%

Table 2: **The percentage of queries that have above or equal to the median, or worst performance (in terms of NDCG@10 and ERR@10 of the *SD+W* run) compared to other TREC participants. Note: the numbers do not add up to 100% in each column.**

6.3 Further Analysis

6.3.1 Comparison with the Other Teams

While a full comparison with the other teams is not available at this point, we can look at the distribution of our retrieval scores compared to the **median** and **min** of the overall score distributions. Looking at Table 2, we can state that, overall, the performance of our runs was above the median. On a less positive note, however, there is a considerable number of cases, for which no relevant results are retrieved. We analyze these cases in the next section.

6.3.2 Failure Analysis

In this section we examine the queries, for which no relevant documents were retrieved at the top thirty ranks. Overall, there are eight such queries in the evaluated set. These queries are:

```
discovery channel store
how to build a fence
income tax return online
to be or not to be that is the question
the wall
kiwi
titan
rice
```

We hypothesize that the poor performance of the three first queries on the list is actually a result of Wikipedia promotion. For these queries, the majority of the top retrieved results are Wikipedia pages, which are non-relevant to these queries. For the next two queries in the above list, stopword removal, which was performed as a part of the indexing process, is the reason that no relevant documents are retrieved. The rest of the poor-performing queries on the list are single-word queries, for which the sequential dependence model (see Section 2) fails to provide any significant benefits. In addition, the pseudo-relevance feedback fails to provide any performance improvements for most of these queries, since no relevant documents are

retrieved by the original non-expanded query. For future submissions, a different, more complex, scoring method should be developed to target such single-word queries.

7 Conclusions

In our submissions for TREC 2010 Web Track, we utilize quality-biased ranking methods that are aimed to promote documents that potentially contain high-quality content, and penalize spam and low-quality documents. Our experiments with the ad hoc web topics from TREC 2010 show that features such as the spamminess of the document (as computed by the Waterloo team [6]) and the readability of the document (modeled fraction of stopwords in the document) are very important for improving the precision at the top ranks. Promotion of the high-quality Wikipedia pages leads to further retrieval performance improvements. In addition, we found that using Wikipedia as a high-quality document collection for query expansion can ameliorate some of the negative effects of performing pseudo-relevance feedback from a noisy web collection such as ClueWeb09.

One weak point of our submission was the failure to deal with queries that contain names or quotes, in which stopwords play an integral part (e.g., the famous "to be or not to be" quote). In addition, our methods did not perform well for many single-word queries, where both the initial retrieval and the retrieval with expanded queries yielded no relevant documents at the top ranks. We intend to explore ways of dealing with such queries in our future submissions.

8 Acknowledgment

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #CNS-0934322, and in part by ARRA NSF IIS-9014442. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

[1] J. Allan, J. Aslam, B. Carterette, V. Pavlu, and E. Kanoulas. Million Query Track 2008 overview. In *Proc. of TREC*, 2008.

[2] J. Allan, M. E. Connell, W. B. Croft, F. F. Feng, D. Fisher, and X. Li. INQUERY and TREC-9. In *Proc. of TREC*, 2000.

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

[4] T. Brants and A. Franz. Web 1T 5-gram Version 1, 2006.

[5] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2004 Terabyte Track. In *Proc. of TREC*, 2004.

[6] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. Apr 2010.

[7] J. He, K. Balog, K. Hofmann, E. Meij, M. de Rijke, M. Tsagkias, and W. Weerkamp. Heuristic ranking and diversification of web documents. In *Proc. of TREC*, 2009.

[8] M. Y. Ivory and M. A. Hearst. Improving web site design. *IEEE Internet Computing*, 6(2):56–63, March 2002.

[9] T. Kanungo and D. Orr. Predicting the readability of short web summaries. In *Proc. of WSDM*, New York, NY, USA, 2009.

[10] R. Krovetz. Viewing morphology as an inference process. In *Proc. of SIGIR*, pages 191–203, 1993.

[11] V. Lavrenko and W. B. Croft. Relevance models in information retrieval. Number 13 in Information Retrieval Book Series, pages 11–56. Kluwer, 2003.

[12] J. Lin, D. Metzler, T. Elsayed, and L. Wang. Of ivory and smurfs: Loxodontan mapreduce experiments for web search. In *Proc. of TREC*, 2009.

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

[14] A. Ntoulas and M. Manasse. Detecting spam web pages through content analysis. In *Proc. of WWW*, pages 83–92, 2006.

[15] M. D. Smucker, C. Clarke, and G. V. Cormack. Experiments with ClueWeb09: Relevance feedback and web tracks. In *Proc. of TREC*, 2009.

[16] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. *Proceedings of the International Conference on Intelligence Analysis*, 2004.

[17] Y. Zhou and W. B. Croft. Document quality models for web ad hoc retrieval. In *Proc. of CIKM*, pages 331–332, 2005.