University of Amsterdam at the TREC 2011 Web Track

Marijn Koolen[1] **Jaap Kamps**[1,2]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam

Abstract: In this paper, we document our efforts in participating to the TREC 2011 Web Tracks. We had multiple aims: This year, tougher topics were selected for the Web Track, for which there is less popularity information available. We look at the relative value of anchor text for these less popular topics, and at impact of spam priors. Full-text retrieval on the ClueWeb09 B collection suffers from text spam, especially in the top 5 ranks. The spam prior largely reduces the impact of spam, leading to a boost in precision. We find that, in contrast to the more common queries of last year, anchor text does improve ad hoc retrieval performance of a full-text baseline for less common queries. However, for diversity, mixing anchor text and full-text leads to an improvement. Closer analysis reveals that mixing anchor text and full-text, fewer relevant nuggets are retrieved which cover more subtopics. Anchor text is an effective way of reducing redundancy and increasing coverage of subtopics at the same time.

1 Introduction

The challenge of the Web Track this year is to provide diverse search results for tougher, less popular queries. Therefore, we expect the relevant pages to be fewer in number, as well as less popular than pages targeted by popular queries. This suggests anchor text is less useful. We study the relative value of full-text and anchor text representation. Last year we discovered that spam in category B is mainly affecting full-text runs [6], while anchor-text and popularity measures like PageRank are much less affected. We experiments with different spam re-ranking methods. This year we also experiment with using feedback, which might be more effective for tough topics than popularity-based methods. We use no diversity-specific features.

The rest of this paper is organised as follows. We first describe our experimental setup in Section 2. We discuss our results in Section 3 and provide a more detailed analysis in Section 4. We summarise our findings in Section 5.

2 Experimental Setup

For the Web Track, we experiment with full-text and anchor text representations and a mixture of these two, based on the ClueWeb09 category B collection. We used Indri [3] for indexing, with stopwords removed and terms are stemmed using the Krovetz stemmer. We built the following indexes:

Full-text B: contains document text of all documents in ClueWeb category B.

Title B: field in the Full-text B index, contains the titles of all documents in ClueWeb category B.

Anchor B: contains the anchor text of all documents in ClueWeb category B. All anchors are combined in a bag of words. 37,882,935 documents (75% of all documents) have anchor text and therefore at least one incoming link.

For all runs, we use either Dirichlet smoothing ($\mu = 2500$) or Jelinek-Mercer (JM) smoothing. In Indri, JM smoothing is impletemend as follows:

$$P_{JM}(r|d) = \frac{(1-\lambda) \cdot tf_{r,d}}{|d|} + \lambda \cdot P(r|D) \qquad (1)$$

where d is a document in collection D. We use little smoothing ($\lambda = 0.05$), which was found to be very effective for large collections [4, 5].

For ad hoc search, pages with more text have a higher prior probability of being relevant [7]. Because some web pages have very little textual content, we use a linear document length prior $\beta = 1$. That is, the score of each retrieved document is multiplied by $P(d)$:

$$P_{dl}(d) = \frac{|d|^\beta}{\sum_{d' \in D} |d'|^\beta} \qquad (2)$$

To combat spam, we use the Fusion spam scores provided by Cormack et al. [1]. We turn the spam scores into a spam prior probability and reduce the impact of spam pages by multiplying the retrieval scores by the spam percentile. The retrieval score is combined with either or both priors by multiplying the probabilities:

$$S_L(d) = P_{dl}(d) \cdot P(r|d) \qquad (3)$$

$$S_S(d) = S_{spam}(d) \cdot P(r|d) \tag{4}$$

$$S_{LS}(d) = S_{spam}(d) \cdot P_L(d) \cdot P(r|d) \tag{5}$$

$$\tag{6}$$

where $S_{spam}(d)$ is the spam percentile for d and $P(r|d)$ is either $P_{JM}(r|d)$ (JM smoothing) or $P_{Dir}(r|d)$ (Dirichlet).

Using a length prior on the anchor text representation of documents has an interesting effect, as the length of the anchor text is correlated to the incoming link degree of a page. The anchor text of a link typically consists of one or a few words. The more links a page receives, the more anchor text it has. Therefore, the length prior on the anchor text index promotes web pages that have a large number of incoming links and thus the more important pages.

2.1 Official Runs

We submitted six runs for the Adhoc and Diversity Tasks:

UAmsAnc05LS: Anchor-text run with linear smoothing and linear length and spam priors.

UAmsM705FLS: Mixture of an Anchor run and a Full-text with feedback, with a spam prior used on both runs.

UAmsM705tFLS: Mixture of an Anchor run and a Full-text+Title with feedback, with a spam prior used on both runs.

UAmsM705tiLS: Mixture of Anchor run and Full-text+Title, with linear smoothing and a spam prior used on both runs.

UAmsM7DirExS: Mixture of 70% Full-text+Title and 30% Anchor runs, with Dirichlet smoothing ($\mu = 2500$).

UAmsT05FLS: Full-text run with linear length prior, feedback and a spam prior.

All mixture runs are made by taking 70% of the Full-text score and 30% of the Anchor score.

3 Results

3.1 Ad hoc

Results for the Ad hoc task are shown in Table 1.

Indexes We compare the various indexes (Anchor, Full-text, Full-text+Title and Mix(title)) using JM smoothing and the length and spam priors. The Anchor index is more effective than the Full-text index, but less effective than the Full-text+Title index and the Mix(title) index. Putting more weight on title words improves results of the full-text index (compare $\textit{Full-text}_{LS,JM}$ and $\textit{Full-text} + \textit{Title}_{LS,JM}$). The Mix(title) run is not as effective as the Full-text+Title run, showing that anchor text does not contribute positively to full-text search for ad hoc search. The Anchor run is greatly improved by the length prior, suggesting that the popular pages (which have more incoming links, thus a longer anchor text representation) have a higher probability of being relevant than less popular pages.

Table 1: Results for the 2010 Ad hoc task. Best scores are in boldface. The first 6 runs are the official runs. Runs starting with * are alternative names of the official runs.

	nDCG		**ERR**	
Run id	**@10**	**@20**	**@10**	**@20**
UAmsAnc05LS	0.172	0.156	0.096	0.101
UAmsM705FLS	0.204	0.182	0.106	0.112
UAmsM705tFLS	0.213	0.189	0.108	0.114
UAmsM705tiLS	**0.225**	**0.202**	**0.114**	**0.119**
UAmsM7DirExS	0.155	0.138	0.095	0.100
UAmsT05FLS	0.139	0.152	0.074	0.082
$\textit{Full-text}_{LS,JM}$	0.165	0.171	0.088	0.096
$*\textit{Full-text}_{FLS,JM}$	0.139	0.152	0.074	0.082
$\textit{Full-text} + \textit{Title}_{Dir}$	0.190	0.177	0.087	0.095
$\textit{Full-text} + \textit{Title}_{L,JM}$	0.129	0.140	0.069	0.076
$\textit{Full-text} + \textit{Title}_{LS,JM}$	0.227	**0.217**	0.105	0.113
$\textit{Full-text} + \textit{Title}_{FLS,JM}$	**0.230**	0.210	0.110	0.117
$\textit{Anchor}_{Dir}$	0.115	0.091	0.071	0.074
$\textit{Anchor}_{L,JM}$	0.178	0.154	0.097	0.102
$*\textit{Anchor}_{LS,JM}$	0.172	0.156	0.096	0.101
$*\textit{Mix}_{FLS,JM}$	0.204	0.182	0.106	0.112
$\textit{Mix}_{Dir}$	0.163	0.145	0.093	0.099
$*\textit{Mix}_{S^2,Dir}$	0.155	0.138	0.095	0.100
$*\textit{Mix}(\textit{title})_{LS,JM}$	0.225	0.202	**0.114**	**0.119**
$*\textit{Mix}(\textit{title})_{FLS,JM}$	0.213	0.189	0.108	0.114

Feedback Feedback hurts Full-text run with length and spam prior (compare $\textit{Full-text}_{LS,JM}$ and $\textit{Full-text}_{FLS,JM}$) and the $\textit{Mix}(\textit{title})_{LS,JM}$ run. However, it is effective when more weight is put on the title words ($\textit{Full-text} + \textit{Title}_{LS,JM}$ and $\textit{Full-text} + \textit{Title}_{FLS,JM}$). Perhaps the bare Full-text index has not enough relevance in the top ranks to derive useful feedback terms.

Spam The spam prior is very effective for the Full-text+Title index, but has almost no effect on the Anchor index. Like the more popular queries last year [6], anchor text for these tougher queries seems to be unaffected by spam. Spam is mainly a problem for full-text search.

3.2 Diversity

For the Diversity Tasks we report the official nERR-IA (normalised intent-aware expected reciprocal rank) and α-nDCG measures, and S-recall (subtopic recall) in Table 2. The nERR-IA measure uses collection-dependent normalisation.

Indexes We see the same pattern as for the Ad hoc task. The $\textit{Anchor}_{LS,JM}$ run outperforms the $\textit{Full-text}_{LS,JM}$ run but not the $\textit{Full-text} + \textit{Title}_{LS,JM}$ run. However, for diversity, the Anchor index contributes positively to the Mixture run, making the mixture model more effective than the

Table 2: Impact of length prior on Diversity performance of baseline runs. Best scores are in boldface.

	nERR-IA		α-nDCG		P-IA		S-recall	
Run	10	20	10	20	10	20	10	20
UAmsAnc05LS	0.400	0.409	0.426	0.455	0.208	0.167	0.601	0.695
UAmsM705FLS	0.451	0.457	0.482	0.502	0.253	0.196	0.675	0.713
UAmsM705tFLS	**0.477**	**0.482**	0.505	0.522	0.259	0.205	0.691	0.723
UAmsM705tiLS	0.473	0.479	**0.511**	**0.530**	**0.265**	0.220	**0.723**	**0.745**
UAmsM7DirExS	0.355	0.365	0.387	0.417	0.186	0.152	0.580	0.665
UAmsT05FLS	0.321	0.332	0.366	0.400	0.247	**0.240**	0.554	0.616
$Full\text{-}text_{LS,JM}$	0.359	0.367	0.411	0.436	0.255	0.244	0.644	0.669
$*Full\text{-}text_{FLS,JM}$	0.321	0.332	0.366	0.400	0.247	0.240	0.554	0.616
$Full\text{-}text + Title_{Dir}$	0.367	0.381	0.413	0.456	0.256	0.230	0.606	0.730
$Full\text{-}text + Title_{L,JM}$	0.266	0.279	0.321	0.362	0.200	0.210	0.545	0.651
$Full\text{-}text + Title_{LS,JM}$	0.439	0.447	0.479	0.503	0.276	0.240	0.661	0.710
$Full\text{-}text + Title_{FLS,JM}$	0.450	0.455	0.486	0.504	**0.311**	**0.258**	0.650	0.684
$Anchor_{Dir}$	0.284	0.292	0.306	0.328	0.126	0.090	0.493	0.558
$Anchor_{L,JM}$	0.386	0.393	0.420	0.443	0.210	0.163	0.625	0.675
$*Anchor_{LS,JM}$	0.400	0.409	0.426	0.455	0.208	0.167	0.601	0.695
$*Mix_{FLS,JM}$	0.451	0.457	0.482	0.502	0.253	0.196	0.675	0.713
Mix_{Dir}	0.355	0.364	0.394	0.422	0.190	0.164	0.626	0.675
$*Mix_{S^2,Dir}$	0.355	0.365	0.387	0.417	0.186	0.152	0.580	0.665
$*Mix(Title)_{LS,JM}$	0.473	0.479	**0.511**	**0.530**	0.265	0.220	**0.723**	**0.745**
$*Mix(Title)_{FLS,JM}$	**0.477**	**0.482**	0.505	0.522	0.259	0.205	0.691	0.723

Full-text+Title index. For tough topics, anchor text is more effective for diversity than for ad hoc search.

Feeback

The big difference in P-IA@10 between $Full\text{-}text + Title_{LS,JM}$ and $Full\text{-}text + Title_{FLS,JM}$ suggests that feedback is good for diversity when applied to a relatively good full-text baseline. However, on the mixture runs (bottom 2 rows of Table 2), feedback is not effective.

Spam

As we saw for the Ad hoc task, the diversity of the full-text runs is improved substantially by using spam priors. The spam priors affect the Anchor runs in an interesting way. As we already noted last year, the spam scores not only indicate spamminess of documents, but also different quality aspects. The spam scores can improve results lists that have no spam to start with. Here, we see that the spam prior helps ERR-IA and α-nDCG at both cutoffs, but P-IA and S-recall only at rank 20. The extreme spam prior (spam percentile squared) has little impact on diversity.

4 Analysis

In this section, we perform a further analysis of the results and look for reasons why the anchor text in category B is more effective than the anchor text in category A. We also look at the impact of spam on the performance of our runs. This year, judged documents were labelled as being either irrelevant, relevant, a key resource, a home page targeted by the query or junk/spam. We analyse our runs using these labels.

Table 3: Statistics on the TREC 2010 Ad Hoc assessments over categories A and B

Description	**2011**	**2010**
Total	19,381	25,329
Spam	1019	1431
Non-rel.	15,205	18,665
Relevant	2038	4018
Key	711	1077
Nav	408	138
Rel+Key+Nav	3157	5233

We first look at the relevance assessments themselves. In Table 3 we compare the Ad hoc relevance judgements of this year and last year. Clearly, the tougher topics result in a lower number of relevant documents. Yet there is a larger number of navigational pages this year. This is somewhat surprising given that for tougher topics there is less incentive to use popularity-based measures, which are well-known techniques for navigational search [2, 7].

4.1 Spam

Next, we look at the percentage of results in the top 20 that are labeled as spam (Figure 1). All the official runs use a spam prior and have relatively little spam in the top ranks. We compare them against two Full-text+Title runs that use no spam prior. These latter two runs suffer from spam mainly in the highest ranks, with 34–42% of the top 1 results being spam documents. The Full-text+Title run with

Figure 1: Percentage of results that are labeled spam

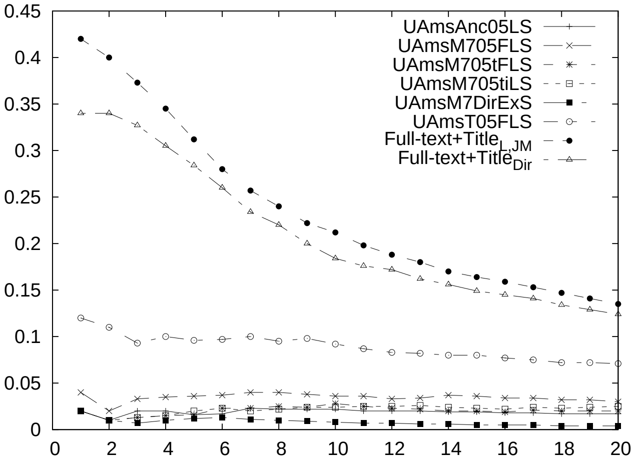
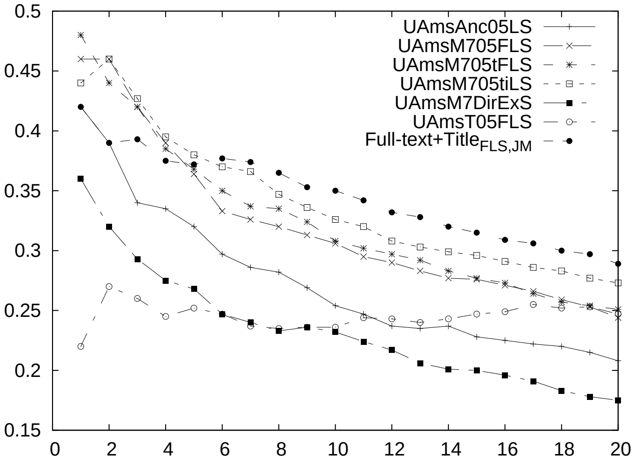
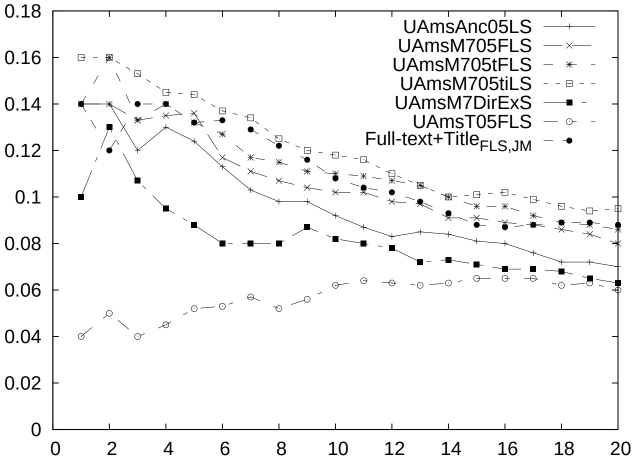


Figure 2: Percentage of results that are labeled relevant, highly relevant or navigational



Dirichlet smoothing suffers less from spam than the one with JM smoothing and a length prior. It seems that spam documents are more common among long documents, which suggests that spammers stuff documents with large amounts of keywords. Among the official runs, the pure full-text run UAmsT05FLS suffers more from spam than the other runs, even though it makes use of the spam priors. The runs on the Anchor index (not shown here) have almost no spam, which is another indicator that spam is mainly a problem for full-text retrieval.

Figure 3: Percentage of results that are labeled as key resource



4.2 Relevance

In Figure 2 we look at the percentage of results labeled as relevant (including key resources and navigational target pages). The official mixture runs with length and spam priors (the 3 runs starting UAmsM705) have a high precision (0.45–0.50) in the first few ranks, which slowly drops to around 0.25–0.28. The full-text run $Full\text{-}text + Title_{FLS,JM}$ starts with lower precision, but keeps it more stable and overtakes the mixture runs around rank 5 or 6. The anchor text helps for very early precision (up to rank 5), but after that reduces the quality of the results with respect to full-text retrieval.

If we look at the percentage of results labeled as key resource (Figure 3), we see that the Mixture model retrieves more key resources than the Anchor model, which might be simply because it retrieves more relevant pages (thus has a higher precision in general). The $Full\text{-}text + Title_{FLS,JM}$ run is close to the best mixture models, but remains below them. Compared to the cumulative relevance in Figure 2, the Anchor text is more effective for identifying key resources.

The percentage of results labeled as navigational target is shown in Figure 4. Most of the official runs have a very small number of navigational pages in the top 20 results. Surprisingly, the UAmsT05FLS run, which performs well below the other official runs on the official evaluation measures (Tables 1 and 2), has the most navigational targets in the top 20. Given the established effectiveness of anchor text for navigational search [2], we would expect the Anchor and Mix runs to find more navigational pages than the plain Full-text index.

4.3 Diversity and Multi-faceted Documents

We saw that the anchor text index contributes positively to the mixture model for diversity but not for ad hoc search. In other words, it reduces the number of relevant documents retrieved in the top ranks, but increases the coverage of multiple subtopics. Is this because the anchor text helps finding documents that cover different subtopics (thereby reducing redundancy) or because it helps finding documents that cover multiple subtopics (retrieving more relevant nuggets with fewer documents).

Figure 4: Percentage of results that are labeled as navigational target

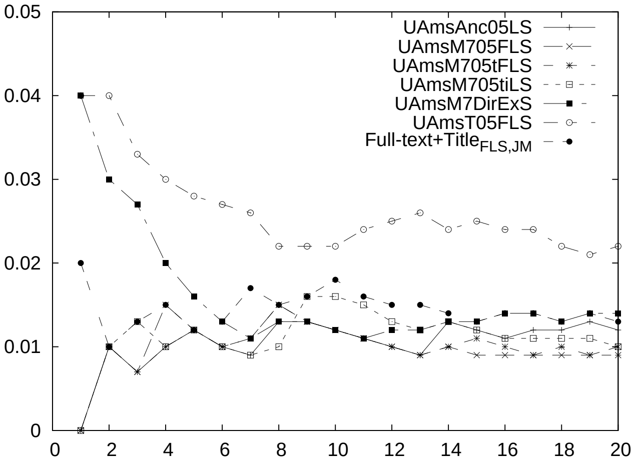


Table 4: Mean (median) number of relevant documents, multi-faceted documents and relevant nuggets in the top 10

Run	**# docs.**	# multi-fac.	# nuggets
$Full\text{-}text + Title_{LS,JM}$	5.08 (5)	2.68 (2)	9.08 (7)
$Anchor_{LS,JM}$	3.82 (3)	1.92 (1)	6.78 (4)
$Mix(title)_{LS,JM}$	4.84 (5)	2.62 (2)	8.70 (7)

The first subtopic is the same as the overall topic, and is a general intent. Many documents that are relevant to other, more specific subtopics are also relevant to the general subtopic. As a consequence, there are many multi-faceted documents, i.e., documents covering more than one subtopic. These documents give a high gain, making it important for systems to return multi-faceted documents.

In Table 4 we see the mean (median) number of relevant documents, multi-faceted documents and relevant nuggets in the top 10 results of the Full-text+Title, Anchor and Mix runs, all with JM smoothing and length and spam priors. The results in this table show that the Mix run has fewer relevant documents and nuggets than the Full-text+Title run, but as shown in Tables 1 and 2, outperforms the Full-text+Title run for diversity. With fewer relevant nuggets, this must mean the Mix run is less redundant than the Full-text+Title run. The anchor text representation selects documents covering different subtopics from those selected by the full-text representation.

The full-text run has a mean of 5.08 relevant documents in the top 10, and 9.08 relevant nuggets (1.79 nuggets per relevant document). With just over half of the relevant documents being multi-faceted, this means the multi-faceted documents often cover more than 2 subtopics. With an average of 3.28 subtopics per topic, this means full coverage can often be attained with one or two relevant documents, which suggests focusing on multi-faceted documents is important for good performance on the official evaluation measures. We will look at the impact of multi-faceted documents more closely in future work.

5 Conclusions

In this paper, we detailed our official runs for the TREC 2011 Web Track and performed an initial analysis of the results. We now summarise our preliminary findings.

With this years tough topics, anchor text is not effective for ad hoc search when compared to a full-text baseline which puts more weight on query terms occurring in the title. The mixture of anchor text and full-text does not lead to an improvement in early precision. For diversity, however, anchor text can contribute positively to the mixture model, by bringing relevant documents in the top ranks that cover different subtopics from the documents retrieved by the full-text index.

Feedback can increase precision of the full-text index, but does not improve diversity (in terms of subtopic recall) for either the full-text index or the mixture model.

Using spam indicators is very effective for both ad hoc retrieval and diversity, as the full-text index suffers severely from text spam. We saw this with the more popular queries in the 2010 Web Track as well. The anchor text representation is less targeted by spammers.

In future work we will look more closely at the difference between anchor text and full-text retrieval, and the impact of multi-faceted documents

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, grant # 612.066.513 and 639.072.601

References

[1] 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.

[2] N. Craswell, D. Hawking, and S. E. Robertson. Effective site finding using link anchor information. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *SIGIR*, pages 250–257. ACM, 2001. ISBN 1-58113-331-6.

[3] Indri. Language modeling meets inference networks, 2009. http://www.lemurproject.org/indri/.

[4] J. Kamps. Effective smoothing for a terabyte of text. In E. M. Voorhees and L. P. Buckland, editors, *The Fourteenth Text REtrieval Conference (TREC 2005)*. National Institute of Standards and Technology. NIST Special Publication 500-266, 2006.

[5] J. Kamps. Experiments with document and query representations for a terabyte of text. In E. M. Voorhees and L. P. Buckland, editors, *The Fifteenth Text REtrieval Conference (TREC 2006)*. National Institute of Standards and Technology. NIST Special Publication 500-272, 2007.

[6] J. Kamps, R. Kaptein, and M. Koolen. Using anchor text, spam filtering and Wikipedia for web search and entity ranking. In E. M. Voorhees and L. P. Buckland, editors, *The Ninteenth Text REtrieval Conference Proceedings (TREC 2010)*. National Institute for Standards and Technology, 2011.

[7] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, New York NY, USA, 2002.