Reducing Redundancy with Anchor Text and Spam Priors

Koolen, M.; Kamps, J.

Published in:
The Twentieth Text REtrieval Conference Proceedings (TREC 2011)

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Reducing Redundancy with Anchor Text and Spam Priors

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 implemented as follows:

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

(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 $P(d)$:

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

(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:

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

(3)
\[ S_S(d) = S_{spam}(d) \cdot P(r|d) \]  
\[ S_{LS}(d) = S_{spam}(d) \cdot P_L(d) \cdot P(r|d) \]

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 run with feedback, with a spam prior used on both runs.
- **UamsM705tFLS**: Mixture of an Anchor run and a Full-text+Title run with feedback, with a spam prior used on both runs.
- **UamsM705tLS**: 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 Full-text\(_{LS,JM}\) and Full-text + 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.

**Feedback** Feedback hurts Full-text run with length and spam prior (compare Full-text\(_{LS,JM}\) and Full-text\(_{FLS,JM}\)) and the Mix(title)\(_{LS,JM}\) run. However, it is effective when more weight is put on the title words (Full-text + Title\(_{LS,JM}\) and Full-text + 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 DiversityTasks we report the official nERR-IA (normalised intent-aware expected reciprocal rank) and \( \alpha \)-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 Anchor\(_{LS,JM}\) run outperforms the Full-text\(_{LS,JM}\) run but not the Full-text + Title\(_{LS,JM}\) run. However, for diversity, the Anchor index contributes positively to the Mixure 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.

<table>
<thead>
<tr>
<th>Run</th>
<th>nERR-IA 10</th>
<th>nERR-IA 20</th>
<th>α-nDCG 10</th>
<th>α-nDCG 20</th>
<th>P-IA 10</th>
<th>P-IA 20</th>
<th>S-recall 10</th>
<th>S-recall 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>UAmsAnc05LS</td>
<td>0.400</td>
<td>0.409</td>
<td>0.426</td>
<td>0.455</td>
<td>0.208</td>
<td>0.167</td>
<td>0.601</td>
<td>0.695</td>
</tr>
<tr>
<td>UAmsM705FLS</td>
<td>0.451</td>
<td>0.457</td>
<td>0.482</td>
<td>0.502</td>
<td>0.253</td>
<td>0.196</td>
<td>0.675</td>
<td>0.713</td>
</tr>
<tr>
<td>UAmsM705tFLS</td>
<td>0.477</td>
<td>0.482</td>
<td>0.505</td>
<td>0.522</td>
<td>0.259</td>
<td>0.205</td>
<td>0.691</td>
<td>0.723</td>
</tr>
<tr>
<td>UAmsM705FLS</td>
<td>0.473</td>
<td>0.479</td>
<td>0.511</td>
<td>0.530</td>
<td>0.265</td>
<td>0.220</td>
<td>0.723</td>
<td>0.745</td>
</tr>
<tr>
<td>UAmsM705tFLS</td>
<td>0.477</td>
<td>0.482</td>
<td>0.505</td>
<td>0.522</td>
<td>0.259</td>
<td>0.205</td>
<td>0.723</td>
<td>0.745</td>
</tr>
<tr>
<td>UAmsM705tFLS</td>
<td>0.355</td>
<td>0.365</td>
<td>0.387</td>
<td>0.417</td>
<td>0.186</td>
<td>0.152</td>
<td>0.580</td>
<td>0.665</td>
</tr>
<tr>
<td>UAmsM705tFLS</td>
<td>0.321</td>
<td>0.332</td>
<td>0.366</td>
<td>0.400</td>
<td>0.247</td>
<td>0.240</td>
<td>0.723</td>
<td>0.745</td>
</tr>
<tr>
<td>Full-textLS,JM</td>
<td>0.359</td>
<td>0.367</td>
<td>0.411</td>
<td>0.436</td>
<td>0.255</td>
<td>0.244</td>
<td>0.644</td>
<td>0.669</td>
</tr>
<tr>
<td>Full-text+TitleLS,JM</td>
<td>0.321</td>
<td>0.332</td>
<td>0.366</td>
<td>0.400</td>
<td>0.247</td>
<td>0.240</td>
<td>0.554</td>
<td>0.616</td>
</tr>
<tr>
<td>Full-text+TitleFLS,JM</td>
<td>0.367</td>
<td>0.381</td>
<td>0.413</td>
<td>0.456</td>
<td>0.256</td>
<td>0.230</td>
<td>0.606</td>
<td>0.730</td>
</tr>
<tr>
<td>Full-text+TitleLS,JM</td>
<td>0.266</td>
<td>0.279</td>
<td>0.321</td>
<td>0.362</td>
<td>0.200</td>
<td>0.210</td>
<td>0.545</td>
<td>0.651</td>
</tr>
<tr>
<td>Full-text+TitleFLS,JM</td>
<td>0.439</td>
<td>0.447</td>
<td>0.479</td>
<td>0.503</td>
<td>0.276</td>
<td>0.240</td>
<td>0.661</td>
<td>0.710</td>
</tr>
<tr>
<td>Full-text+TitleFLS,JM</td>
<td>0.450</td>
<td>0.455</td>
<td>0.486</td>
<td>0.504</td>
<td>0.311</td>
<td>0.258</td>
<td>0.650</td>
<td>0.684</td>
</tr>
<tr>
<td>AnchorDir</td>
<td>0.284</td>
<td>0.292</td>
<td>0.306</td>
<td>0.328</td>
<td>0.126</td>
<td>0.090</td>
<td>0.493</td>
<td>0.558</td>
</tr>
<tr>
<td>AnchorLS,JM</td>
<td>0.386</td>
<td>0.393</td>
<td>0.420</td>
<td>0.443</td>
<td>0.210</td>
<td>0.163</td>
<td>0.625</td>
<td>0.675</td>
</tr>
<tr>
<td>AnchorLS,JM</td>
<td>0.400</td>
<td>0.409</td>
<td>0.426</td>
<td>0.455</td>
<td>0.208</td>
<td>0.167</td>
<td>0.601</td>
<td>0.695</td>
</tr>
<tr>
<td>MixFLS,JM</td>
<td>0.451</td>
<td>0.457</td>
<td>0.482</td>
<td>0.502</td>
<td>0.253</td>
<td>0.196</td>
<td>0.675</td>
<td>0.713</td>
</tr>
<tr>
<td>MixFLS,JM</td>
<td>0.355</td>
<td>0.364</td>
<td>0.394</td>
<td>0.422</td>
<td>0.190</td>
<td>0.164</td>
<td>0.626</td>
<td>0.675</td>
</tr>
<tr>
<td>Mix(Title)LS,JM</td>
<td>0.473</td>
<td>0.479</td>
<td>0.511</td>
<td>0.530</td>
<td>0.265</td>
<td>0.220</td>
<td>0.723</td>
<td>0.745</td>
</tr>
<tr>
<td>Mix(Title)FLS,JM</td>
<td>0.477</td>
<td>0.482</td>
<td>0.505</td>
<td>0.522</td>
<td>0.259</td>
<td>0.205</td>
<td>0.691</td>
<td>0.723</td>
</tr>
</tbody>
</table>

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

**Feedback**

The big difference in P-IA@10 between Full-text + TitleLS,JM and Full-text + TitleFLS,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

<table>
<thead>
<tr>
<th>Description</th>
<th>2011</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>19,381</td>
<td>25,329</td>
</tr>
<tr>
<td>Spam</td>
<td>1019</td>
<td>1431</td>
</tr>
<tr>
<td>Non-rel.</td>
<td>15,205</td>
<td>18,665</td>
</tr>
<tr>
<td>Relevant</td>
<td>2038</td>
<td>4018</td>
</tr>
<tr>
<td>Key</td>
<td>711</td>
<td>1077</td>
</tr>
<tr>
<td>Nav</td>
<td>408</td>
<td>138</td>
</tr>
<tr>
<td>Rel+Key+Nav</td>
<td>3157</td>
<td>5233</td>
</tr>
</tbody>
</table>

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
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.

4.2 Relevance

In Figure 2 we look at the percentage of results labeled relevant, highly relevant or navigational. 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 + TitleFLS,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 + TitleFLS,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 Subtopic Recall

How many subtopics are covered by each run if we consider the whole result lists? In Figure 5 we show the subtopic recall of the official runs up to rank 10,000. As we already saw in Table 2, the UAmsT05FLS run has the highest subtopic recall at rank 10, but the UAmsT05FLS run overtakes all others at rank 100 and achieves almost total recall (0.995) at rank 1000. At rank 10,000, most runs achieve close to total recall. The anchor text run UAmsAnc05LS remains
more or less stable between ranks 1000 and 10,000, with a
subtopic recall of around 0.93. This figure shows that sys-
tems based on full-text retrieval return results to cover (al-
most) all subtopics and have the potential to substantially
improve subtopic recall in the early ranks by removing re-
dundancy (relevant and non-relevant).

4.4 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 mul-
tiple subtopics. Is this because the anchor text helps find-
ing documents that cover different subtopics (thereby reduc-
ing redundancy) or because it helps finding documents that
cover multiple subtopics (retrieving more relevant nuggets
with fewer documents).

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 gen-
eral 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 im-
portant 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 rep-
resentation.

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 of-
ten 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.

<table>
<thead>
<tr>
<th>Run</th>
<th># docs.</th>
<th># multi-fac.</th>
<th># nuggets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-text + TitleLS,JM</td>
<td>5.08 (5)</td>
<td>2.68 (2)</td>
<td>9.08 (7)</td>
</tr>
<tr>
<td>AnchorLS,JM</td>
<td>3.82 (3)</td>
<td>1.92 (1)</td>
<td>6.78 (4)</td>
</tr>
<tr>
<td>Mix(title)LS,JM</td>
<td>4.84 (5)</td>
<td>2.62 (2)</td>
<td>8.70 (7)</td>
</tr>
</tbody>
</table>

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 im-
provement in early precision. For diversity, however, anchor
text can contribute positively to the mixture model, by bring-
ing relevant documents in the top ranks that cover different
subtopics from the documents retrieved by the full-text in-
dex.
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.

Systems based on full-text retrieval attain a high subtopic recall further down the result list, showing that full-text retrieval has the potential to improve diversity in the top ranks by removing relevant and non-relevant redundancy. Combining full-text with anchor text is an effective way to decrease both types of redundancy.

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


