Using Anchor Text, Spam Filtering and Wikipedia for Web Search and Entity Ranking

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

[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 2010 Entity Ranking and Web Tracks. We had multiple aims: For the Web Track we wanted to compare the effectiveness of anchor text of the category A and B collections and the impact of global document quality measures such as PageRank and spam scores. For the Entity Ranking Track, we use Wikipedia as a pivot to find relevant entities on the Web. We find that documents in ClueWeb09 category B have a higher probability of being retrieved than other documents in category A. In ClueWeb09 category B, spam is mainly an issue for full-text retrieval. Anchor text suffers little from spam. Spam scores can be used to filter spam but also to find key resources. Documents that are least likely to be spam tend to be high-quality results.

1 Introduction

For the Web Track, we experiment with three anchor text variants from two indexes. One index contains all the incoming anchor text of the category A collection, the other index contains only the incoming anchor text of the category B collection. A third variant is derived from the category A index, where we filter on the category B results, to see if the extra anchor text for category B pages, from category A pages, improves the effectiveness. Further, we experiment with combining the retrieval score with PageRank scores and spam classification scores, and filtering results based on spam scores.

Our approach to the TREC Entity Ranking track is similar to the approach we took last year [6]. We adjusted our approach to fit in with the new result format, and to include the ClueWeb Category A document collection. The TREC entity ranking track investigates the problem of related entity finding, where entity types are limited to people, organisations and products. We approach this task as an Entity Ranking task by not using the given input entity. Also we do not use the general entity types of people, organisations and products, instead we have manually assigned more specific target entity types which are also Wikipedia categories. To retrieve entities within Wikipedia, we exploit the category information which has been proven to work for this task. To find the corresponding web entity home pages, we follow the external links on the Wikipedia pages, and search an anchor text index for the page title.

This paper consists of two parts. In the first part, in Section 2, we discuss our experiments for the Web Track. The second part details our Entity Ranking experiments in Section 3. We summarise our findings in Section 4.

2 Web Track

For the Web Track, we experiment with incoming anchor text representation based on either the category A or category B collections.

2.1 Experimental Set-up

For the Web Track runs we used Indri [3] for indexing, with stopwords removed and terms are stemmed using the Krovetz stemmer. We built the following indexes:

Text B: contains document text 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.

Anchor A: contains the anchor text of all documents in the English part of ClueWeb category A, kindly provided by the University of Twente [2]. In total 440,678,986 documents (87% of all English documents) have anchor text. There are 45,077,244 category B documents within this set (90% of all category B documents). We finished our index for category A after the official submission deadline, so we have no official runs based on this index.

For all runs, we use Jelinek-Mercer smoothing, which is implemented in Indri as follows:

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

where d is a document in collection D. We use little smoothing ($\lambda = 0.1$), 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 [8]. 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 $P(d)$:

$$P(d) = \frac{|d|^\beta}{\sum_{d' \in D} |d'|^\beta} \quad (2)$$

The final retrieval score S_{ret} is computed as:

$$S_{ret} = P(d) \cdot P(r|d) \quad (3)$$

.

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.

We used the PageRank scores computed over the entire category A collection provided by CMU.[1] To combat spam, we use the Fusion spam scores provided by Cormack et al. [1]. These spam scores are percentiles based on the log-odds that a page is spam. Documents in the lower percentiles are most likely to be spam, while documents in the higher percentiles are least likely to be spam.

2.2 Official Runs

We look at the impact of filtering spam pages and re-ranking retrieval results by multiplying the retrieval scores by either the PageRank score or the spam percentile. This is computed as:

$$S_{PR}(d) = PR(d) \cdot S_{ret}(d) \quad (4)$$
$$S_{SR}(d) = Spam(d) \cdot S_{ret}(d) \quad (5)$$

where $S_{ret}(d)$ is the Indri retrieval score for document d, $PR(d)$ is the PageRank score for d and $Spam(d)$ is the spam percentile for d.

We submitted three runs for the Adhoc Task:

UAMSA10d2a8: Mixture of document and anchor-text runs of the category B indexes, with a linear length prior probability for both document and anchor-text representations. Scores are combined 0.2 document score + 0.8 anchor-text score.

UAMSA10mSF30: Combination of category B document and anchor-text runs with linear length priors for document and anchor-text representations. Scores combined as 0.2 document score + 0.8 anchor-text score. Results are post-filtered on spam using the Waterloo spam rankings, thresholded at the 30% spammiest pages.

UAMSA10mSFPR: Mixture of category B document and anchor-text runs with linear length priors on document and anchor-text representation. The mixture run scores are multiplied by the CMU PageRank scores and spam-filtered using the Waterloo Fusion spam percentiles, thresholded at the 30% spammiest pages.

We submitted three runs for the Diversity Task:

UAMSD10ancB: Anchor-text run with linear length prior on anchor-text representation using category B.

UAMSD10ancPR: Category B anchor-text run with linear length prior on the anchor-text representation. Retrieval scores are multiplied by the CMU PageRank scores.

UAMSD10aSRfu: Category B anchor-text run with linear length prior on the anchor-text representation. Retrieval scores are multiplied by the Fusion spam percentiles.

2.3 Results

Results for the Ad hoc task are shown in Table 1. We include a number of unofficial runs for further analysis. We make the following observations:

- Of the official runs, the baseline mixture run UAMSA10d2a8 has the highest MAP. Document quality indicators do not help average precision. However, the spam filter (UAMSA10mSF30) is effective for early precision.
- The official UAMSA10mSFPR run performs very poorly, because of a error in the multiplication of the retrieval and PageRank scores.
- The anchor text only run UAMSD10ancB is very similar to the mixture run. This is probably because, in the mixture run, the anchor text score dominates the full-text score. When we combine the anchor text score with the spam percentiles (UAMSD10aSRfu), early precision increases. The spam scores are effective for ad hoc search. This is further discussed in Section 2.4. The PageRank scores (UAMSD10ancPR) are ineffective for the anchor text run. This is not due to any error as with the UAMSA10mSFPR run.

[1] See: http://boston.lti.cs.cmu.edu/clueweb09/wiki/tiki-index.php?page=PageRank.

Table 1: Results for the 2010 Ad hoc task. Best scores are in boldface.

Run id	MAP	MRR	nDCG@10
UAMSA10d2a8	**0.0486**	0.4504	0.1906
UAMSA10mSF30	0.0473	0.4709	0.1949
UAMSA10mSFPR	0.0029	0.0498	0.0113
UAMSD10aSRfu	0.0455	**0.5244**	**0.2053**
UAMSD10ancB	0.0465	0.4494	0.1906
UAMSD10ancPR	0.0250	0.2027	0.0666
Mix B, length	0.0486	0.**4504**	**0.1906**
Anchor B, length	0.0465	0.4494	**0.1906**
Text B, length	**0.0871**	0.2160	0.1108
Anchor A, length	0.0274	0.3376	0.1052
Anchor A, filter B, length	0.0294	0.3720	0.1172

- As expected, the full-text run Text B has a higher MAP than the anchor text and mixture runs. While it has lower early precision, it finds many more relevant documents. Note that the Text B run was not submitted, and therefore has a substantial number of unjudged results in the top ranks; precision is probably underestimated.

- The category B anchor text index is more effective than the category A anchor text index. Although performance of the category A index improves when we filter out pages that do not occur in category B, it falls behind performance of the category B anchor text run. If we filter out the results that are not in category B, the results improve. It seems that the documents in category B have a higher probability of being relevant. We will further analyse this difference in the next section.

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

The performance of the mixture run UAMSA10d2a8 and the anchor text run UAMSD10ancB are similar. Again, this is probably due to high weight on the anchor text score in the mixture model. Filtering out pages below the 30th percentile (UAMSA10mSF30) has a small positive effect. Re-ranking the results by combining the anchor text score with the spam percentile (UAMSD10aSRfu) leads to bigger improvements at rank 5. Combining the anchor text run with PageRank (UAMSD10ancPR) hurts diversity performance.

The anchor text run (Anchor B) is clearly more diverse than the full-text run (Text B). But, because some of the top results of the Text B run are unjudged, these scores are a lower bound. The mixture run (Mix B) leads to a small improvement in nERR-IA@5 and strec@5.

The anchor text index of category A has lower scores than the Anchor B run. If we filter on category B, the scores go up, again suggesting that the category B documents are of higher quality.

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

Description	Category A	Category B
Documents	500M	50M (10%)
Judgements	18,161	11,189 (62%)
Spam	655	301 (46%)
Irrelevant	13,217	8,429 (64%)
Relevant	3,329	1880 (56%)
Key	833	499 (60%)
Nav	127	80 (63%)

2.4 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.

We first look at the relevance assessments themselves, in Table 3. The category B part of ClueWeb09 is a 10% subset of category A. In total, 18,161 query-document pairs were judged, the majority of which are for documents in the category B collection. The top 100 results of the official runs seem to have mainly category B documents. This could be due to many participants submitting category B-only runs, or because documents in category B are ranked higher than the rest of the documents in category A. Of the pages judged as spam, only 46% comes from category B. This suggests that category B contains less spam. The relevant pages (including key resources and navigational pages) are as frequent in the judged documents of category B as in the judged documents of category A.

If we look at the top 100 results of the Anchor A run, we find that 53% of the results are category B documents. This shows that, at least for anchor text, the category B documents are more often retrieved than non-category B documents in category A. But why does the Anchor B run perform so much better than the Anchor A, even when we filter the Anchor A run on category B? In Figure 1 we look at the percentage of non-judged results in the top 100. Because the Anchor B run is an official submission, the top 20 results are judged. For the other two runs, many of the top results or not judged, which, at least partially, explains why the Anchor A runs are score lower than the Anchor B run.

In the rest of this section, we look at the official submissions. Next, we look at the percentage of results in the top 20 that are labeled as spam (Figure 2).[2] Only the Text B run suffers from spam, and especially at the highest ranks,

[2]Because of the error with the UAMSA10mSFPR run and the poor performance of the UAMSD10ancPR run, we leave these runs out of our analysis, to keep the figures easy to read.

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

	nERR-IA			nNRBP	strec@		
Run id	5	10	20		5	10	20
UAMSA10d2a8	0.232	0.248	0.264	0.228	0.345	**0.460**	**0.591**
UAMSA10mSF30	0.238	**0.252**	**0.268**	**0.234**	0.340	0.437	0.568
UAMSA10mSFPR	0.018	0.024	0.029	0.019	0.031	0.076	0.129
UAMSD10aSRfu	**0.241**	**0.252**	0.267	0.232	**0.367**	0.455	0.570
UAMSD10ancB	0.230	0.248	0.264	0.228	0.331	**0.460**	**0.591**
UAMSD10ancPR	0.089	0.102	0.116	0.084	0.189	0.284	0.435
Text B	0.089	0.107	0.125	0.074	0.194	0.301	0.448
Anchor B length	0.230	**0.248**	**0.264**	**0.228**	0.331	**0.460**	**0.591**
Mix B	**0.232**	**0.248**	**0.264**	**0.228**	**0.345**	**0.460**	**0.591**
Anchor A length	0.168	0.176	0.186	0.161	0.304	0.359	0.475
Anchor A, filter B, length	0.190	0.203	0.214	0.193	0.259	0.346	0.456

Figure 1: Percentage of results that are not judged

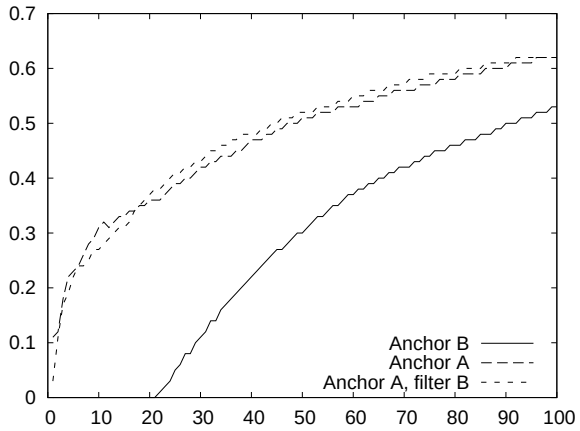


Figure 2: Percentage of results that are labeled spam

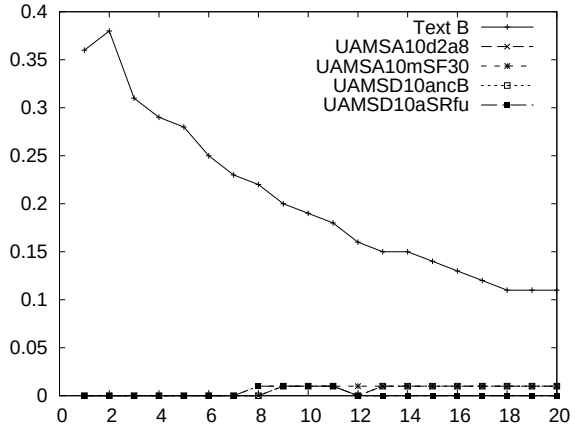
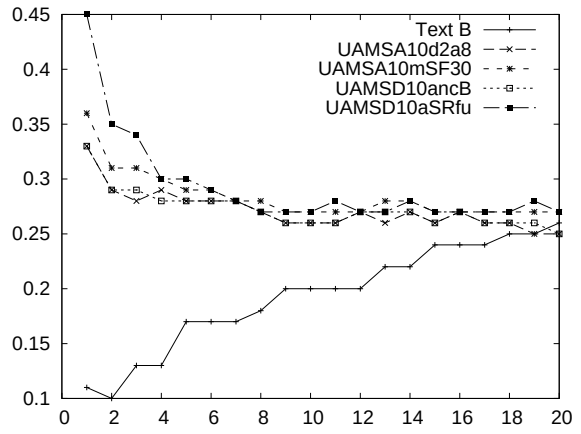


Figure 3: Percentage of results that are labeled relevant



with 36% of the results at rank 1 being spam. At rank 2 the percentage is even higher (38%). At lower ranks, the percentage gradually drops to around 10%. All other runs, which are mainly based on the anchor text index, do not suffer from spam. At least in category B, anchor text seems not to be abused by spammers.

In Figure 3 we look at the percentage of results labeled as relevant (including key resources and navigational target pages). Here we see that the precision of the Text B run increases with rank, which is probably due to the fact that the amount of spam at each rank gradually drops with increasing rank. Note that not all of the Text B results in the top 20 are judged (from 6% at rank 1 up to 18% at rank 20), so the actual percentage of relevant documents might be higher (as well as the percentage of spam). Of the official runs, the ranking based on both the anchor text score and the spam percentile (UAMSD10aSRfu) has the highest precision at rank 1. However, at rank 4 and beyond, the official runs have a very similar precision. Also, precision remains relatively stable after rank 4.

If we look at the percentage of results labeled as key re-

Figure 4: Percentage of results that are labeled as key resource

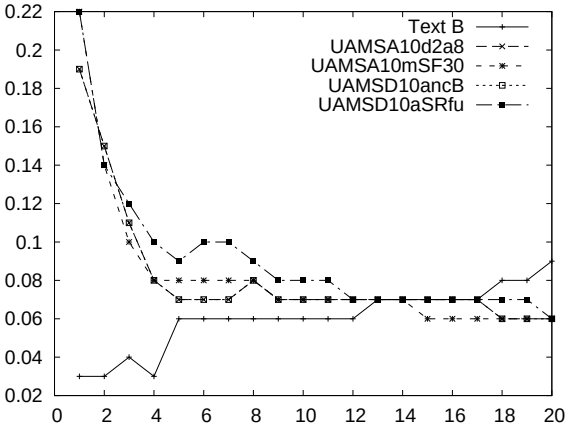
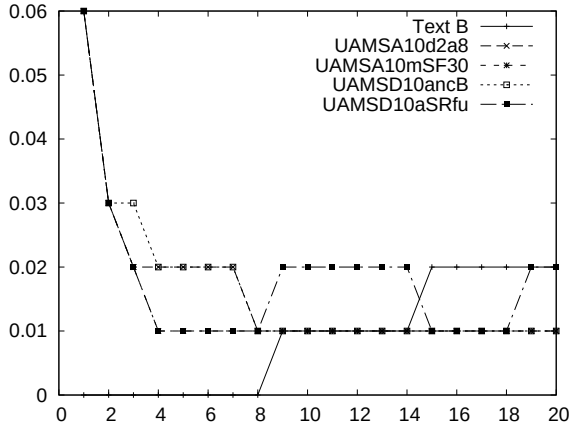


Figure 5: Percentage of results that are labeled as navigational target



source (Figure 4), we see again that the UAMSD10aSRfu run has a slightly higher percentage at rank 1—22% as opposed to 19% of the other 3 official runs—but the percentage rapidly drops to around 8% for these runs. If we promote documents that are least likely to be spam, we find more key resources in the top of the ranking. This shows that the spam scores not only indicate whether a document is spam or not, but provide an overall indicator of document quality as well. The Text B run has a low percentage at rank 1 (again, possibly due to spam), but catches up with the anchor text based runs at rank 13 and from rank 18 even outperforms them. This is in line with the higher MAP of the Text B run; beyond the first ranks, its precision is better than that of the anchor text and mixture runs.

The percentage of results labeled as navigational target is shown in Figure 5. The Text B run finds no navigational targets before rank 9, whereas the official runs start with 6% navigational targets at rank 1. However, this percentage quickly drops to between 1 and 2 percent. As with the key resources, anchor text easily finds one or a few highly linked home page and other important pages.

3 Entity Ranking

For the entity ranking track, we have experimented with different approaches, which are discussed in this section. To complete the task of entity ranking, we split the task up into three steps:

1. Rank all Wikipedia pages according to their match to the narrative from the query topic.
2. Rerank the top retrieved Wikipedia pages, according to their match with the target entity types
3. Find home pages belonging to the retrieved Wikipedia pages

3.1 Retrieving entities in Wikipedia

Our approach exploits the category information in Wikipedia. The target entity types which are assigned during topic creation (people, organisations, products and locations) are too general for our purposes. Instead we have assigned manually more specific entity types to each query. These entity types can also be assigned automatically by pseudo-relevance feedback, i.e. take the top N results from the initial ranking created in step 1 of the entity ranking process, and assign the most frequently occurring category as the target entity type.

Our initial run is a language model run with a document length prior created with Indri [3]. To rerank the pages according to their match with the target entity types, we use the following algorithm. KL-divergence is used to calculate distances between categories, and calculate a category score that is high when the distance is small, and the categories are similar as follows:

$$S_{cat}(C_d|C_t) = -\sum_{t \in D} \left(P(t|C_t) * \log \left(\frac{P(t|C_t)}{P(t|C_d)} \right) \right) \quad (6)$$

where d is a document, i.e. an answer entity, C_t is a target category and C_d a category assigned to a document. The score for an answer entity in relation to a target category $S(d|C_t)$ is the highest score, or shortest distance from any of the document categories to the target category. A linear combination of the initial score as calculated in step 1 and the category score produces the final score by which the Wikipedia pages are ranked.

3.2 Retrieving home pages for Wikipedia Entities

In the third and last step of our approach we retrieve home pages associated with the retrieved Wikipedia pages. In the

Wikipedia context we consider each Wikipedia page as an entity. The Wikipedia page title is the label or name of the entity. We experiment with three methods to find Web pages associated with Wikipedia pages:

1. External links: Follow the links in the External links section of the Wikipedia page.
2. Anchor text: Take the Wikipedia page title as query, and retrieve pages from an anchor text index using a length prior.
3. Combined: When no external link is available search the anchor text.

For each Wikipedia page we only include the first result of the associated Web pages. In the 'External Links' method results are skipped if no external links exist in the document collection for the Wikipedia result.

3.3 Runs

Since there are no results available at the time of writing, we can not report on the results. The following runs have been submitted:

- UAcatscombB : based on the Wikipedia run using category information, web pages are retrieved from ClueWeb category B using the combined method.
- UAcatslinkA: based on the Wikipedia run using category information, web pages are retrieved from ClueWeb category A using the external links.
- UAbaseanchB: based on the initial Wikipedia run without using category information, web pages are retrieved from Clueweb category B using the anchor text method.
- UAbaselinkA: based on the initial Wikipedia run without using category information, web pages are retrieved from ClueWeb category A using the external links.

4 Conclusions

In this paper, we detailed our official runs for the TREC 2010 Web Track and Entity Ranking Track and performed an initial analysis of the results. We now summarise our preliminary findings.

For the Web Track, we wanted to compare the anchor text representations of ClueWeb09 category A and category B and look at the impact of spam scores.

The larger category A anchor text index covers many more documents than the category B anchor text index. It also increases the coverage and amount of anchor text of category B documents. However, the category B anchor text run outperforms the category A anchor text run, even if we filter the latter to retain only the category B results.

Our analysis of the relevance judgements shows that the majority of the Ad hoc judgements are for documents in category B, but relatively fewer of the documents labeled as spam are in category B. This shows that the top results of the official runs consist mainly of category B documents and also suggests that documents in category B are of higher quality than other documents in ClueWeb09. We also found that the category A anchor text run mainly has category B documents in the top 100 results, which suggests that category B documents have a higher probability of being retrieved. Another explanation for the lower scores of the category A anchor text run is that it has many non-judged results in the top ranks, so the evaluation scores might be underestimated.

In our experiments, only the full-text index suffers from spam, indicating that anchor text is less targeted by spammers. The Fusion spam scores can help reduce spam, but also used as indicators of document quality. If we rerank search results by combining the retrieval score with the spam score, we can improve the effectiveness of anchor text—which, in our experiments, does not suffer from spam—for locating key resources.

For the Entity Ranking Track, we experimented with using Wikipedia as a pivot to find related entities in the larger Web. However, no results are available at the time of writing. We will conduct detailed experiments and analysis when results are provided.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, grant # 612.066.513, 639.072.601, and 640.001.501).

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] D. Hiemstra and C. Hauff. MIREX: MapReduce Information Retrieval Experiments. Technical Report TR-CTIT-10-15, 2010. ISSN 1381-3625. http://eprints.eemcs.utwente.nl/17797/.

[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] R. Kaptein, M. Koolen, and J. Kamps. Result diversity and entity ranking experiments: Text, anchors, links, and wikipedia. In E. M. Voorhees and L. P. Buckland, editors, *The Eighteenth Text REtrieval Conference Proceedings (TREC 2009)*. National Institute for Standards and Technology. NIST Special Publication, 2010.

[7] M. Koolen and J. Kamps. The importance of anchor-text for ad hoc search revisited. In H.-H. Chen, E. N. Efthimiadis, J. Savoy, F. Crestani, and S. Marchand-Maillet, editors, *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129. ACM Press, New York NY, USA, 2010.

[8] 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.