What is the Importance of Anchor Text for Ad Hoc Search?

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

[1] Archives and Information Studies, University of Amsterdam, The Netherlands
[2] ISLA, Informatics Institute, University of Amsterdam, The Netherlands
{m.h.a.koolen,kamps}@uva.nl

ABSTRACT

It is generally believed that propagated anchor text is very important for effective Web search, but many years of TREC Web retrieval research failed to establish the effectiveness of link evidence for ad hoc retrieval on Web collections. In this paper we use the new TREC 2009 Web Track collection to study the impact of collection size and link density on the effectiveness of anchor-text for Web ad hoc retrieval. Our main findings are that anchor-text outperforms full-text retrieval in terms of early precision and an improvement in overall precision when combined with it. Other findings are that, contrary to expectations, link density has little impact on effectiveness, while the size of the collection has a substantial impact on the quantity, quality and effectiveness of anchor text. This paper is based on [6].

1. INTRODUCTION

The use of anchor text for Web retrieval is well studied, with the broad conclusion that it is very effective for finding entry pages of sites–often outperforming approaches based on document text alone–but not for ad hoc search. Some speculated that the number of (inter-server) links in the TREC collections was too low and that the collections might be too small for anchors to be effective [3]. Others pointed at the difference between traditional ad hoc retrieval studied at TREC and actual Web search. Web searchers tend to "prefer the entry page of a well-known topical site to an isolated piece of text, no matter how relevant" [4]. Although the switch to more Web-centric search tasks like home page and named page finding showed link information to be very effective [2, 7], there is no clear explanation of why anchor text is not effective for ad hoc retrieval. To study the value of link information, Gurrin and Smeaton [3] suggested a representative test-collection needs to be sufficiently large and have sufficiently high inter- and intra-server link densities. At the TREC 2009 Web Track [1] a new, large Web collection—ClueWeb09—was introduced, which is much larger than previous collections and was crawled to reflect Tier 1 of a commercial search engine, so has a relatively dense link structure, urging us to revisit the question:

- What is the importance of anchor text for ad hoc search?

2. INITIAL EXPERIMENTS

We indexed the ClueWeb09 category B, which is a 50 million pages subset of the full ClueWeb09, using Indri with Krovetz stemming and stopword removal. We created two indexes, a *full-text* index and an *anchor text* index containing only the propagated anchor text of ClueWeb09 B. The full-text and anchor text runs use the Indri language model approach and linear smoothing with $\lambda_{collection} = 0.15$. Documents are scored using the document length as a prior probability $p(d) = \frac{|d|}{|D|}$, where d is a document in collection D. We also made a mixture run, combining the full-text and anchor runs using the weighting $S_{mix}(d) = 0.7 \cdot S_{full}(d) + 0.3 \cdot S_{anchor}(d)$.

Table 1: Results for the 2009 Adhoc Task. Significant differences ($p > 0.95$, denoted $^\circ$) are with respect to the full text run

Run	Full collection statMAP	Full collection MPC(30)	No Wikipedia statMAP	No Wikipedia MPC(30)
Text	0.1442	0.3079	0.1038	0.2557
Anchor	0.0567	**0.5558**	0.0617	0.4289
Mix	**0.1643**$^\circ$	0.4812°	**0.1213**	**0.4773**
Text · In-degree	0.1098	0.2694	0.0746	0.2059
UDWAxQEWeb	0.1999	0.5010	–	–
uogTrdphCEwP	0.2072	0.4966	–	–
ICTNETADRun4	0.1746	0.4368	–	–

2.1 Results

The results are shown in Table 1. We test for significant changes with respect to the full-text baseline using a one-tailed bootstrap test with 100,000 resamples. The *Anchor* run has a low statMAP compared to the *Text* run. A possible explanation is that many pages in the collection have no or few incoming links, including many relevant pages. In contrast, anchor text is effective for early precision. The *Anchor* run scores better on MPC(30) than the *Text* run and supports the above explanation for its low statMAP score. More importantly, the *Mix* run leads to significant improvements in statMAP showing that the two indexes are complementary and that Web structure can be used to improve ad hoc search. To put this into perspective, we compared them against the top 3 groups of the TREC 2009 Web Ad hoc task (according to MPC(30), bottom 3 rows). The runs of the top 3 groups score substantially better on statMAP, but lower on MPC(30). This shows that anchor text alone can meet or exceed the precision of the top-performing systems.

Perhaps anchor text is more effective than in previous TREC experiments because this collection contains the full Wikipedia, which has a dense link structure and many anchors matching the titles of the target pages. Columns 4 and 5 in Table 1 show the results of these runs. The *Anchor* run still has higher early precision and the *Mix* run still has higher statMAP than the *Text* run. Wikipedia is not the reason for the effectiveness of anchor text. In sum, this new Web collection finally shows the long expected value of Web link structure for ad hoc search.

DIR'10, 4 Feruary, 2011, Amsterdam, Netherlands.

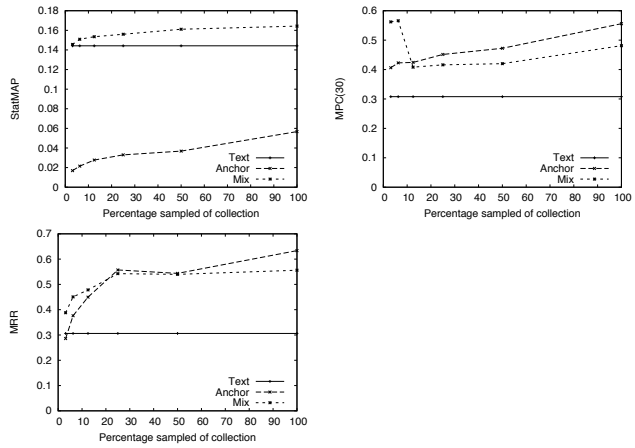

Figure 1: Impact of link sampling on effectiveness of full-text, anchor text and mixture runs.

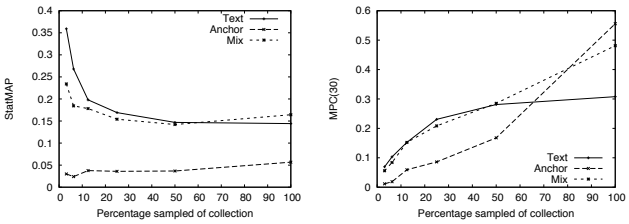

Figure 2: Impact of page sampling on effectiveness of full-text, anchor text and mixture runs.

3. WHY ANCHOR TEXT WORKS

In this section we seek to understand what makes the anchor text representation effective. We look at the impact of link density and collection size, which we do by down-sampling either links or pages.

3.1 The Impact of Link Density

We filter links by randomly selecting n% of all documents and removing their outgoing links. If we randomly sample 50% of the pages and remove the outgoing links of those pages, we would expect to end up with roughly 50% of all the links. The impact of sampling links on the effectiveness of full-text and anchor text is shown in Figure 1. The full-text index is not affected by link sampling, hence the straight line in the figures. The statMAP (top left) of the *Anchor* run slowly decreases as we remove more links because the index covers fewer pages. The *Mix* run scores better at statMAP with even the smallest samples of links, indicating that even very few links can improve the *Text* run. The MPC(30) scores (top right) of the *Anchor* run stay well above the *Text* score. We note that below 12.5% of the links (less than 3 incoming links per page), the density is well below the link densities of earlier TREC Web collections. The impact of link density seems small. To rule out that the MPC(30) score is over-estimated we transformed the relevance judgements to traditional binary judgements and looked at the Mean Reciprocal Rank (MRR, bottom left of Figure 1), which cannot over-estimate. It supports that anchor text gives better early precision than full-text. Link density plays a role at low densities, but its impact stabilises quickly.

3.2 The Impact of Collection Size

Next, we look at the impact of the collection size. We randomly remove n% pages from the collection, and thereby lose both the outgoing and incoming links of those pages. Thus, if we sample 50% of the pages, we remove more than 50% of the links. One of the favourable aspects of randomly sampling pages is that the probability of relevance is unaffected [5]. The impact of sampling pages on the effectiveness of full-text and anchor text is shown in Figure 2. The statMAP (left figure) of the *Text* run goes up slowly—possibly due to losing topics with little relevance—while for the *Anchor* run it goes down slowly. The *Text* run gains precision at rank 30 (MPC(30), right figure) as the collections grows, as predicted [5]. The anchor text precision is more affected by collection size. With half the collection, anchor text is nowhere near as effective as full-text. With fewer relevant documents left, and an increasingly smaller coverage of the collection, it becomes harder to find relevant pages through anchor text. For precision at a fixed cut-off, the impact of the collection size is much larger for anchor text than for full-text.

4. CONCLUSIONS

Our main finding is that in contrast with earlier results, the anchor text leads to significant improvements in retrieval effectiveness for ad hoc informational search. Link density has little impact on anchor text effectiveness, while collection size has a big impact on the anchor text representations, affecting quantity, quality and effectiveness. Full-text search is less affected by collection size.

Perhaps the main contribution of this paper is that it solves the apparent contradiction between the experiences of Internet search engines, and the results of experiments at TREC. This turns the earlier negative results into something positive in a sense: they aid to our understanding of when and why link evidence works, and when not.

Acknowledgments

This work was generously supported by the Netherlands Organisation for Scientific Research (NWO, grants # 612.066.513, 639.-072.601, and 640.001.501).

REFERENCES

[1] C. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In *TREC*, 2009.

[2] N. Craswell, D. Hawking, and S. E. Robertson. Effective site finding using link anchor information. In *SIGIR*, pages 250–257, 2001.

[3] C. Gurrin and A. F. Smeaton. Replicating web structure in small-scale test collections. *Inf. Retr.*, 7:239–263, 2004.

[4] D. Hawking and N. Craswell. Very large scale retrieval and web search. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. MIT Press, 2005.

[5] D. Hawking and S. Robertson. On collection size and retrieval effectiveness. *Inf. Retr.*, 6(1):99–105, 2003.

[6] 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.

[7] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR*, pages 27–34. ACM, 2002.