The Importance of Anchor Text for Ad Hoc Search Revisited

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 as offered by the commercial search engines. "Google Bombs" are a notable illustration of this. However, many years of TREC Web retrieval research failed to establish the effectiveness of link evidence for ad hoc retrieval on Web collections. The ultimate resolution to this dilemma was that typical Web search is very different from the traditional ad hoc methodology. So far, however, no one has established why link information, like incoming link degree or anchor text, does not help ad hoc retrieval effectiveness. Several possible explanations were given, including the collections being too small for anchors to be effective, and the density of the link graph being too low.

The new TREC 2009 Web Track collection is substantially larger than previous collections and has a dense link graph. Our main finding is that propagated anchor text outperforms full-text retrieval in terms of early precision, and in combination with it, gives an improvement in overall precision. We then analyse the impact of link density and collection size by down-sampling the number of links and the number of pages respectively.

Other findings are that, contrary to expectations, (inter-server) 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. We also compare the diversity of the search results of anchor text and full-text approaches, which show that anchor text performs significantly better than full-text search and confirm our findings for the ad hoc search task.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms: Experimentation, Measurement, Performance

Keywords: Ad hoc, Anchor text, Collection size, Link density

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. Based on claims from commercial search engine companies, over the course of several years of Web search experiments at TREC [34], organisers and participants have tried to establish the effectiveness of link information, including anchor text, for retrieval. Despite the enthusiasm and effort of many participating groups, in the first two years, 1999–2000, participants failed to show any improvements due to link information [17]. After the first year, the main reason was deemed to be the low number of inter-server links [19]. For the second year, an artificially crafted collection with more inter-server links (WT10g, [4]) was used, and participants combined content information with link evidence such as PageRank, HITS and anchor text, but again without success. This time, the main difference between the results at TREC and the general belief that link information is valuable, was considered to be the difference in search tasks [15, 33]. Typical Web search behaviour is very different from the user model assumed for the traditional ad hoc methodology. Web searchers tend to "prefer the entry page of a well-known topical site to an isolated piece of text, no matter how relevant" [17]. According to Hawking and Craswell [17, p.215]:

> Hyperlink and other web evidence is highly valuable for some types of search task, but not for others.

Although the switch to more Web-centric search tasks like home page and named page finding showed link information to be very effective for these tasks [9, 24, 26, 29], there is no clear explanation of why anchor text is not effective for ad hoc retrieval. Anchor text provides short summaries often by different authors about the topic of a page. This could potentially improve the document representation of web pages that have incoming links, and thereby the precision of search results. On the Web, which is infinitely large, early precision is the important criterion. Recall is almost impossible to measure, but also not important for most users.

Gurrin and Smeaton [14] pointed out that the inter-server link density of the WT10g collection was still very low, and extracted a subset of the collection, *WT-dense*, which has a much higher inter-server link density. Within this tiny subset they found that a combination of content and link information could improve precision on the ad hoc topics of the TREC-9 Web track. This led them to come up with a list of requirements a representative test-collection must satisfy to study the value of link information. A good Web collection needs to be sufficiently large and have sufficiently high inter- and intra-server link densities.

The size issue was addressed in the Terabyte Tracks of 2004–2006, which used the GOV2 collection, based on a crawl of the `.gov` domain in 2004, consisting of 25 million documents.[1] Again, anchor text was found to be highly effective for Web-centric tasks,

[1]The crawl on the `.gov` domain was exhausted before reaching the targeted 100 million pages, and plans to rectify this by crawling additional pages from the `.edu` domain were never realised.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

but not for ad hoc search [22, 23]. However, the .gov domain is very different in nature from the .com domain used for the crawl on which the WT10g collection is based, and the .GOV2 collection has fewer incoming links per page. Thus, although it is larger than the earlier Web Track collections, its link density is much lower, making it hard to investigate the impact of collection size.

At the TREC 2009 Web Track [6] a new, large Web collection—ClueWeb09 [7]—was introduced and the traditional Ad hoc Task was paired with the new Diversity task. This new collection is much larger than the collections used at TREC 8 and 9, and was crawled to reflect Tier 1 of a commercial search engine, so should have a relatively dense link structure, allowing us to study both aspects of collection size and link density. If a large number of documents and a high link density are indeed requirements for anchor text to be effective, this new collection might finally reveal us its potential. This urges us to revisit the question:

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

Surely, the issue of having enough (inter-server) links for anchor text is critical for its success on any search task, but perhaps the link density needs to be higher for anchor text to be effective for ad hoc retrieval than for entry page finding. Intuitively, the number of links in the collection plays a direct role in the quantity of anchor text and might therefore affect its quality as well. Links within the same site are often navigational links, with anchor terms such as 'click', 'here' and 'next' [11]. Therefore, it is generally assumed that links between sites are more meaningful, including their anchor text [27].

The other factor mentioned by Gurrin and Smeaton [14] is the size of the collection. We know that precision increases with collection size [18], which holds for document text indexes, but should hold for anchor text indexes as well. With larger collections, the number of documents that have incoming links increases and as a consequence, so does the number of those documents relevant to a given topic. The new ClueWeb collection used at the TREC 2009 Web Track is much larger than the collections used at TREC 8 and 9 and should have many incoming links per page, resulting in more anchor text and thereby possibly better document representations.

The direct relation between anchor text and link in-degree, and thereby the relation between anchor text and the popularity or importance of a page, helps locate home pages and popular pages described by the search terms. This offers an interesting perspective on the task of finding diverse search results. If anchor text is effective for ad hoc search as well, it has the potential to find good results for both informational and navigational information needs.

These considerations lead us to break down our main research question into several, more specific research questions:

- Is anchor text effective for improving ad hoc retrieval?
- What is the impact of link density on the effectiveness of anchor text? And what is the relative importance of inter- and intra-server links?
- What is the impact of collection size on the effectiveness of anchor text?
- What is the importance of anchor text for the diversity of ad hoc search results? And what is the impact of anchor text on informational and navigational information needs?

The rest of the paper is organised as follows. We will first discuss related work in Section 2. In Section 3 we describe our initial experiments to compare the effectiveness of anchor text and full-text search. Then, in Section 4 we analyse how link density and collection size affect the quantity and quality of anchor text, followed by Section 5 where we investigate the diversity of anchor text search results. We draw conclusions in Section 6.

2. RELATED WORK

The importance of anchor text has been studied extensively at TREC. At TREC 8, participants could not show consistent improvements over content-only baselines using link information [19]. This unexpected result led many to believe that the collection had too few inter-server links for link evidence to be effective, in response to which a new collection was constructed focusing on inter-server link density [4]. At the TREC 9 Web Track, the first reported attempts at exploiting anchor text for ad hoc retrieval did not show any improvements either [15]. Singhal and Kaszkiel [32] raised doubts about the TREC evaluation methodology used to model Web search, as they found different results for anchor text when comparing TREC results against their in-house tests.

Several studies [8, 33] pointed at the differences between traditional ad hoc search (as evaluated at TREC) and Web search behaviour. As new, more realistic Web tasks were introduced [16], the value of link information was finally shown [9, 26] and anchor text was found to be be very effective for site- and home page finding tasks. Craswell et al. [10] recently showed the effectiveness of anchor text for diversity. Eiron and McCurley [11] showed that anchor text behaves much like user queries. If Web authors use the same labels to describe pages as Web searchers use to find pages, anchor text can potentially bridge the gap between queries and pages and lead to high precision, if the anchors and pages in the collection are of high quality.

The quality is another important difference between the new ClueWeb collection and previous TREC Web collections, which is related to the way it is constructed, and which directly affects the density of (inter-server) links. Several studies have looked at the impact of crawling policy on the quality [3] and search effectiveness [12, 13] of the crawled collection. Page importance metrics can be used to schedule the most important or useful pages to be crawled first. Since page importance is usually derived using link-based measures such as PageRank [30] or On-line Page Importance Computation [OPIC, 1], which give a higher score to a page if it has more incoming links, the first part of a crawl based on such policies tends to have a high link density. One of the primary goals of creating the ClueWeb data set was "to approximate Tier 1 of a web search engine index" [5]. The category B data set, which we use here, consists of the first 50 million English pages of this crawl.

3. INITIAL EXPERIMENTS

We will first describe our initial experiment with plain full-text and anchor text approaches to see if the relative effectiveness of anchor text merits further investigation.

3.1 Data, Index and Runs

We use the ClueWeb09 category B, which contains a sample of 50 million English pages of the larger ClueWeb09 crawl [7]. As the pages were crawled based on a large set of seed URLs with high PageRank and at later stages were crawled in order of OPIC value, we assume this data set contains many of the most important Web pages and a relatively dense Web graph.

We used Indri [21] for indexing. Stopwords are removed and all other terms are stemmed with the Krovetz stemmer. We created two indexes, a *full-text* index containing only the document text and an *anchor text* index containing only the propagated anchor text.

With anchor text we mean the underlined text to which a hyperlink is anchored. We only use the string of text appearing in the anchor text. To extract anchor text from the ClueWeb09 category B collection, we used the *harvestlinks* method, which comes with Indri. Because *harvestlinks* does not compress its data during processing, which would use up more disk space than we have avail-

able, we only used the harvesting option to extract the anchor text with source url, document ID and target url per bundle of pages, compressed the output, then mapped the target url to a document ID ourselves. We extracted over 1.5 billion links pointing to pages within the collection, with anchor text for more than 75% of all the pages. Quite a large number of pages have multiple links to the same target URL (repeated links). If we collapse those repeated links we end up with 1.18 billion links between just over 50 million pages, which leads to a mean in-degree of 23.30. The median in-degree is 2. For anchor text, repeated links mean more (and potentially different) descriptions from the same source page.

The full-text and anchor text runs use the Indri language model approach and linear smoothing with $\lambda_{collection} = 0.15$. For ad hoc search, the length of a document is related with the probability of relevance. Documents are scored using the document length as a prior probability $p(d) = \frac{|d|}{|D|}$, where d is a document in collection D. The length prior on the anchor text is determined by the total length of the all the anchor text of a particular page. This means that pages with many incoming links have long document representations, while pages with a single incoming link have a very short document representation. The length prior boosts pages with many links, i.e. pages with high in-degrees. We report on the effectiveness of the length prior in [25]. We also made a mixture run, combining the full-text and anchor runs using the weighting $Score_{mix}(d) = 0.7 \cdot Score_{full}(d) + 0.3 \cdot Score_{anchor}(d)$, where the scores are normalised by the sum of the top 1000 scores before addition. If a document d is retrieved by only one index, it receives a zero probability score for the other index. The mixture run was submitted as an official run at the TREC 2009 Web Adhoc task and contributed to the pool. The other two runs can be considered as baseline runs. No tuning was done after the relevance judgements were made available. Furthermore, we have created two runs using the incoming link degree of the pages returned by the *Text* run. We chose to use in-degree as it was to found to be as effective as PageRank [2, 28] but easier to compute. For the *In-degree* run, the full-text results are ranked only by in-degree. For the *Text · In-degree* run we use the in-degree as another document prior in the language model of Indri.

3.2 Results

The results are shown in Table 1. To put this into perspective, we compared them against the results of the best official submissions of other participants. Both MTC and statAP were used to construct the judgement pools. We use statAP [35], as it is more robust when evaluating runs that did not contribute to the pool. We test for significant changes with respect to the full-text baseline using a one-tailed bootstrap test with 100,000 resamples.

Is anchor text effective for ad hoc retrieval? 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. With no or only a few words as document representation, these pages are hard to find using anchor text only, which has consequences for average precision. In contrast, anchor text is very effective for early precision. The *Anchor* run scores much better on MPC(30) than the *Text* run and supports the above explanation for its low statMAP score. The anchor text run ranks the relevant pages in its index highly, but seems to miss many relevant pages. We will further investigate this issue below. 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.

The in-degree priors only hurt the underlying *Text* run. This is not surprising given that in-degree is blind to the topic of a query.

Table 1: Results for the 2009 Adhoc Task. Significance tests are with respect to the full text run, confidence levels are 0.95 ($^\circ$), 0.99 ($^\circ$) and 0.999 ($^\bullet$)

	Full collection		No Wikipedia	
Run	**statMAP**	**MPC(30)**	**statMAP**	**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
In-degree	0.0823	0.1876	0.0592	0.1258
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	–	–

Within a much larger collection containing spam pages and other pages with low PageRank, degree-based link ranking algorithms have an important function of separating the high quality pages from the rest. Within this particular collection, which is crawled to reflect Tier 1, most pages are of high quality, so the work of finding important and reliable pages is already done. Further use of in-degrees only disrupts the subtle relevance ranking of text-based retrieval models. Anchor text gives more precise results because it focuses on the subset of links that have query terms in the anchors, and is thus more sensitive to the topical context than in-degree.

The best runs of the top 3 groups of the TREC 2009 Web Ad hoc task, according to MPC(30), 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.

Over the top 1000, the overlap between the *Anchor* and *Text* runs is 4.6%, showing that anchor text really leads to very different document representations and targets very different pages. The overlap between the Ad hoc judgements and the *Anchor* run is also very small. In the top 5 results of the *Anchor* run, over 30% of the pages are unjudged, while at rank 16, more than half of the results are unjudged. Indeed, it seems that the *Anchor* run is also very different from all runs that contributed to the assessment pool. In contrast, the overlap between the Ad hoc judgements and the *Text* run is much higher. At rank 5, on average less than 11% is unjudged, while at rank 16 just over 22% is unjudged. These percentages are probably much closer to the actual sampling rates. Thus, the improvement of the *Mix* run over the *Anchor* run might simply be caused by the larger number of judged results in that run.

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. Wikipedia pages are edited by many contributors, so the quality might be higher than that of many Web pages. The relevance judgements reveal that more than 21% of the relevant pages in ClueWeb B are Wikipedia pages, while the whole Wikipedia forms only 12% of all the pages in the collection. We built separate full- and anchor text indexes of all non-Wikipedia pages. If the presence of Wikipedia is the main reason for the effectiveness of anchor text, we would expect the non-Wikipedia *Anchor* run to perform worse than the non-Wikipedia *Text* run. Columns 4 and 5 in Table 1 show the results of these runs. Although scores are lower than over the full collection indexes—perhaps partly explaining why ClueWeb B runs tend to outperform ClueWeb A runs [6]—the *Anchor* run still has higher early precision and the *Mix* run still has higher statMAP than the *Text* run. Wikipedia is not the sole reason for the effectiveness of anchor text.

Is anchor text effective for improving ad hoc retrieval? On a large collection of high quality pages, anchor text gives good precision

and in combination with full-text leads to significant improvements in overall precision. This new Web collection finally shows the long expected value of Web link structure for ad hoc search. Gurrin and Smeaton [14] suggested that the benefits of link information for retrieval would become clear with a sufficiently large collection and a high inter-server link density. Our results support their statement and urge us to address these issues.

4. 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. If we down-sample the pages, we can investigate the impact of collection size on the effectiveness of anchor text. If, on the other hand, we keep the number of pages the same, but instead down-sample the links, we can see the impact of link density.

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. If we remove the pages themselves from the collection, we 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. Previous TREC Web collections were smaller which could explain why anchor text was not effective earlier. The number of links does not grow linearly with collection size. How does page sampling affect the link density of the collection?

Randomly sampling pages is different from using earlier stages of the crawl as a smaller collection. The first 25 million pages of the crawl have a different composition from randomly sampling 25 million pages [13]. With random sampling, the most important pages will be affected in the same way as the rest of the pages. If the crawl is stopped at half the number of pages, the collection will still have most of the important pages, as modern crawling strategies focus on crawling the most important pages first. However, since ClueWeb B is assumed to be a subset of Tier 1 of a Web search index, based on a large number of seed URLs, we expect that the composition of any sample of the ClueWeb B collection approximates the composition of the full collection. One of the favourable aspects of randomly sampling pages is that the probability of relevance is unaffected [18].

Down-sampling either pages or links means the anchor text representation of a page changes. For each sample, we have to filter the link graph and anchors, and build a separate anchor text index. For the full-text index, we only have to make separate indexes for the page filtered samples. Sampling links has no impact on the full-text document representations. Since page sampling has an effect on both collection size and link density and link sampling only affects link density, we will first look at the impact of sampling links.

4.1 The Impact of Link Density

If we remove links from the collection, we can expect the performance of anchor text to go down. If we remove all links, the anchor text index is empty and no page will ever be returned. The more links in the collection, the more information we have to distinguish between pages. Therefore, we expect to see effectiveness increase with link density, at least while the density is low. Beyond a certain point the effectiveness might stabilise or become worse.

We filter links by randomly selecting n% of all documents and removing their outgoing links. The impact of sampling outgoing links on the number of inter- and intra-server links is shown in Figure 1. Reading from right to left, both the number of inter- and intra-server links decrease, linear to the sample size. How does this affect the pages with anchor text? If we remove 50% of the links, pages will lose roughly half of their incoming anchors. For

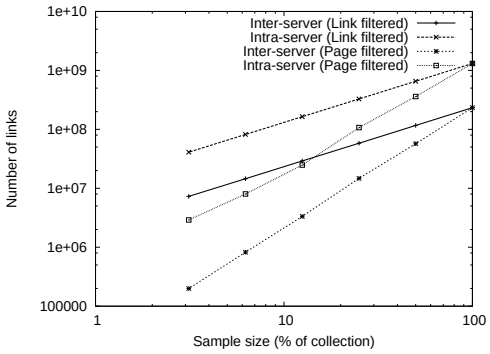

Figure 1: The impact of outgoing link sampling on the number of inter- and intra-server links per sample.

Table 2: Impact of link filtering on the percentage of pages with anchor text

	All pages			Relevant pages		
Percent	**Inter**	**Intra**	**All**	**Inter**	**Intra**	**All**
100.000	15.30	70.26	75.43	25.54	74.46	80.96
50.000	11.41	56.35	61.51	21.04	64.84	71.96
25.000	8.24	43.79	48.36	17.14	54.87	61.97
12.500	5.78	33.06	36.75	13.77	44.15	50.75
6.250	3.94	24.17	26.96	10.94	35.81	41.80
3.125	2.61	17.00	19.00	8.35	28.41	33.33

pages with high in-degree, this might not affect the document representation much. Pages with only one or a few incoming anchors could lose most or all of their anchor text. In other words, the most important pages are robust against random link sampling.

In Table 2 we see how filtering affects the percentage of pages that have at least one incoming link with anchor text. The inter-server links cover only 15% of all pages but 25% of the relevant pages. Apparently, pages with incoming links from other sites have a higher probability of being relevant. The intra-server links cover a much larger part of the collection (70%). The ratio of inter- to intra-server links is about 1 to 5.5. At a sample size of 12.5% the number of intra-server links is lower than the number of inter-server links at a sample size of 100%. However, at 12.5%, the intra-server links cover 33% of the pages (44% of the relevant pages), while at 100% the inter-server links cover only 15% of the pages and 25.54% of the relevant pages. The intra-server links are thus more uniformly distributed, leading to fewer anchors per page. What does this mean for the effectiveness of inter- and intra-server links?

The impact of sampling links on the effectiveness of full-text and anchor text is shown in Figure 2. The full-text index is not affected by link sampling, hence the straight line in the figures. The statMAP 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. We also look at traditional MAP and found very similar results. Contrary to our expectations, with smaller samples, the MPC(30) scores of the anchor text 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. One possible explanation is that the highest quality pages have so many incoming links that they are robust against link sampling. This is reflected in the percentages shown in Table 2. With fewer links, the number

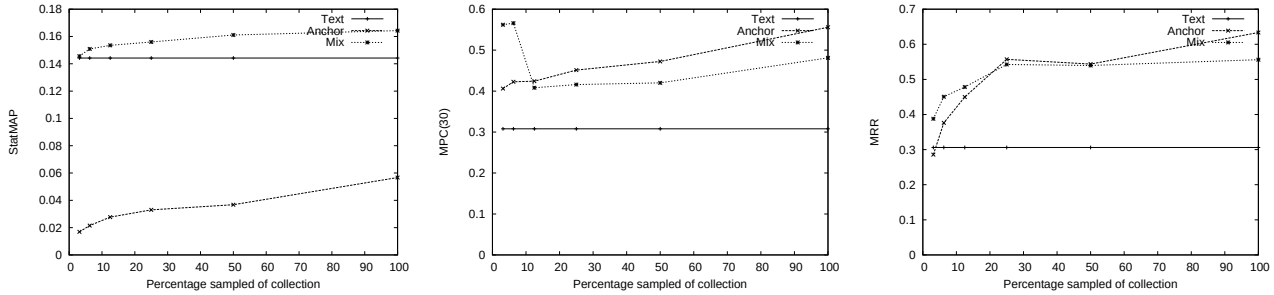

Figure 2: Impact of link sampling on effectiveness of full-text, anchor text and mixture runs.

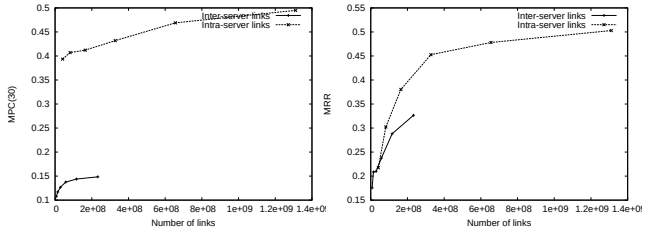

Figure 3: Comparison of inter- and intra-server link anchor effectiveness.

of pages with anchors goes down, but the number of relevant pages decreases more slowly. As the link density goes down, the relevant pages form a larger part of the index. On the other hand, the *Anchor* run might find relevant pages ranked much lower by the runs that contributed the pages to the pool, which represent many estimated relevant pages. 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, right side of Figure 2). The MRR never over-estimates as it simply counts the rank of the highest ranked relevant document. It supports that anchor text gives better early precision than full-text.

What is the qualitative difference between inter- and intra-server links? We already saw there is a big quantitative difference. Most of the links are between pages on the same server (85%). The left side in Figure 3 shows the MPC(30) scores for the inter- and intra-server *Anchor* runs. Of course, any observed difference could be due to the larger quantity of intra-server links. Therefore, we show the scores with the actual number of links on the x-axis. Even at a similar number of links, the intra-server *Anchor* run scores better. An explanation is that the intra-server link anchors cover more pages, because they are more evenly distributed, and can thus find more relevant pages. The difference in MRR between the inter- and intra-server links is smaller, and both scores go up with more links, showing that both are better able to identify relevant pages with higher link density. However, the impact of link density quickly stabilises beyond a certain point. In the first tier of a Web index, containing high quality pages, there seems to be little qualitative difference between inter- and intra-server links. Insight in search engine optimisation may have taught Web site owners to make internal anchor text more meaningful. The larger quantity of intra-server links, and their more even distribution makes them more effective for finding multiple relevant pages.

What is the impact of link density on the effectiveness of anchor text? It plays a role at low densities, but its impact stabilises quickly. Inter- and intra-server links have different distributions and a different coverage of the collection. Within a collection of high PageRank pages, the difference between inter- and intra-server links is more quantitative than qualitative. Without a crawling policy that focuses on finding high quality pages first, the quality of the crawl and therefore the quality of the intra-server links might go down. Inter-server links might be more robust, as they tend to have less navigational anchors and are harder to use nepotistically.

Table 3: Number of documents and topics per sample

Percent	Size in Docs	# Rel. Docs	Topics	# Rel./topic
100.000	50,220,423	4,002	49	81.57
50.000	25,110,211	1,987	49	40.55
25.000	12,555,105	965	47	20.53
12.500	6,277,552	486	46	10.57
6.250	3,138,776	253	44	5.75
3.125	1,569,388	132	42	3.14

4.2 The Impact of Collection Size

Next, we look at the impact of the collection size. How effective is anchor text if we reduce the size of the collection? To see this, we need to down-sample the collection, which we do by randomly selecting half of the pages. We show the number of (relevant) pages in each filtered sample in Table 3. At each step, the ratio of all pages and relevant pages is roughly the same. If we remove 50% of the pages in the collection, and remove those from the relevance judgements as well, we end up with about 50% of the relevant pages. For smaller samples this has consequences for the number of topics with any relevant document. At 3.125%, we have 132 relevance judgements left for 42 topics (3.14 relevant pages per topic). Although the number of topics is still large enough to be representative, the low number of relevant pages per topic might make per topic results unreliable.[2]

As mentioned above, if we sample pages, the number of links does not decrease linearly to the sample size. In Table 4 we see the percentage of pages in the sample that have at least one incoming link. One interesting observation is that page sampling has a similar impact on the coverage of inter-server links as link sampling, but a very different impact on the coverage of intra-server links. If we sample 3.125% of the links (Table 2), the intra-server anchors cover 17% of all pages, while if we sample 3.125% of the pages, the intra-server anchors cover 43% of the remaining pages. Yet

[2]The relevance judgements of the Ad hoc task include the probability of a page being included in the assessment pool. This probability is used to estimate how many pages a pooled page represents, which is based on the size of the collection. If a relevant page has a pool probability of 0.2 it represents $\frac{1}{0.2} = 5$ relevant pages in the full collection. With uniform down-sampling, the probability of a page being sampled stays the same, but the number of pages it represents is proportional to the sample size. At a sample size of 25% of the full collection it represents 25% of the relevant pages it represents in the full collection. Most judged pages have a unit pooling probability, so are not affected by down-sampling.

Table 4: Impact of page filtering on the percentage of pages with anchor text

	All pages			Relevant pages		
Percent	**Inter**	**Intra**	**All**	**Inter**	**Intra**	**All**
100.000	15.30	70.26	75.43	25.54	74.46	80.96
50.000	11.40	61.09	65.49	19.23	67.54	73.63
25.000	8.23	53.96	57.50	15.34	61.04	66.32
12.500	5.77	48.73	51.41	14.40	59.26	63.99
6.250	3.94	45.07	46.98	10.28	52.57	56.52
3.125	2.61	42.59	43.88	7.58	50.76	53.79

the inter-server anchors cover 2.61% of the collection, whether we sample 3.125% of the links or the pages.

The impact of sampling pages on the effectiveness of full-text and anchor text is shown in Figure 4. 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. Theory suggests that statMAP should remain relatively stable over random samples of a collection [18]. The drop in statMAP for the *Anchor* run can be explained by looking at the precision scores. The *Text* run gains precision at rank 30 (MPC(30), centre figure) as the collections grows, as predicted [18]. The anchor text precision is more affected by collection size. With half the collection, anchor text is nowhere near as effective as full-text. The MRR (right figure) of the *Text* run is similar to that of the statMAP. The average rank of the first relevant document increases slowly, partly due to losing topics. However, the MRR of the *Anchor* run drops rapidly with smaller samples. With fewer relevant documents left, and an increasingly smaller coverage of the collection, it becomes harder to find relevant pages through anchor text.

What is the impact of the collection size on 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. First, collection size affects the anchor text representation but not the full-text representation. Second, the number of pages in the full-text index grows linearly with collection size, but more than linearly for anchor text. For ad hoc search, where the task is to find pages with relevant text no matter their popularity, this coverage is essential. We stress again the importance of having a collection of high quality pages; expanding the collection with low quality pages will probably also lower the quality of link anchors.

In summary, link density does not explain the effectiveness of anchor text as even a small number of links lead to improved performance. Intra-server links are at least as effective as inter-server links because they cover a larger part of the collection. The size of the collection plays a larger role than link density, because it has a larger impact on the number of pages with anchor text and the quality of their anchor text representation. The effectiveness of anchor text rapidly increases as we expand the collection. We now look at how these findings hold up in a more Web-oriented search task.

5. DIVERSITY TASK

We now look at the effectiveness of full-text and anchor text approaches for result diversity, where the task is to present a diverse set of results in the top 10 or 20 results. The Diversity task uses the same topics, but breaks them down into a number of informational and navigational sub-topics. This allows a deeper analysis of the various strengths and weaknesses of full-text and anchor text approaches. The task thus combines both ad hoc search and entry and named page finding. Given the earlier successes with anchor text for the latter tasks [9, 29], and its good performance on ad hoc

Table 5: Ad hoc and Diversity evaluation using the Diversity relevance judgements. Significance tests are with respect to the full text run, confidence levels are 0.95 ($^\circ$), 0.99 ($^\circledcirc$) and 0.999 ($^\bullet$)

Run	α-nDCG@10	nDCG@10	IA-P@10	P@10
$Text$	0.120	0.1564	0.054	0.1700
$Anchor$	**0.257**$^\bullet$	**0.2780**$^\bullet$	**0.082**$^\circ$	**0.2460**$^\circ$
$Text + Anchor$	0.223$^\bullet$	0.2459$^\bullet$	0.083$^\circledcirc$	0.2420$^\circledcirc$
uogTrDYCcsB	0.282	–	0.132	–
ICTNETDivR3	0.272	–	0.095	–
UamsDancTFb1	0.250	–	0.079	–

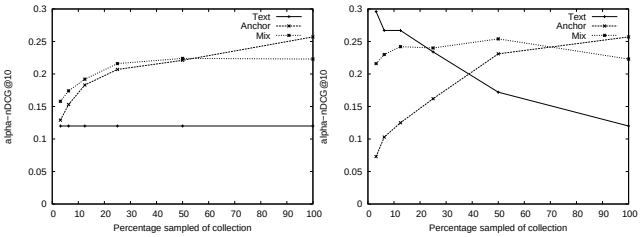

Figure 5: Impact of link sampling (left) and page sampling (right) on diversity of full-text, anchor text and mixture runs.

precision described in the previous sections, we conjecture anchor text to perform better than full-text on the diversity task as well.

Anchor text shares characteristics with queries [11]. A well-known problem with IR research is the fact that different users can type the same query but have different information needs. Thus, a query can and will be used to search for different types of information, or different aspects or facets of a topic. The same might hold for anchor text. Two Web page authors can use the same anchor text to link to different target pages, covering different topics or different aspects of the same topic.The Diversity relevance judgements are based on a pool of top 20 results of all official runs and were made independently from the ad hoc judgements, thus provide a sanity check on our findings from the ad hoc experiments.

We first compare the performance of the anchor text and full-text runs on the full collection in Table 5. The official measures are α-nDCG@10 and IA-P@10. The differences are very clear. Anchor text performs significantly better than full-text search. In the diversity score, the ad hoc relevance ranking plays an important role. If the relevance ranking is low, that is, there are many irrelevant documents in the top ranks, the diversity score will therefore be low as well. Therefore, we turned the diversity qrels into standard TREC qrels by making a page relevant for a topic if it is relevant for at least one sub-topic. This allows us to compare the diversity specific measures with their non-diversity counterparts (columns 3 and 5 in Table 5). We see very similar patterns for α-nDCG@10 and normal nDCG@10. The same holds for IA-P@10 and P@10. These results suggest the anchor text scores much better on the diversity measures simply because it has a better underlying relevance ranking, supporting our findings in the previous section.

What is the impact of link density and collection size on the diversity of anchor text results? We use the same indexes and runs as described in the previous section. The results are shown in Figure 5. Sampling links (left figure) has a similar effect on diversity as on ad hoc search. Again, we see that anchor text becomes more effective with more links, but the improvements become smaller beyond 25% of the links. However, even with 3.125% of the links, the anchor text run is at least as good as the full-text run. Again, link density seems to matter only at very low densities. We looked

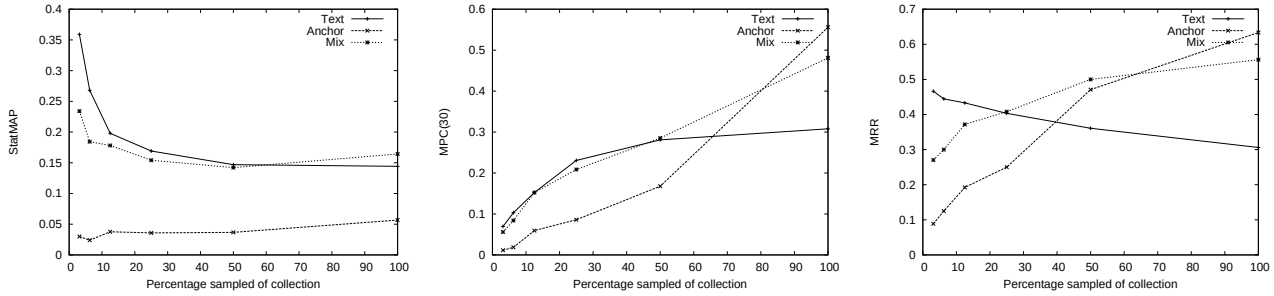

Figure 4: Impact of page sampling on effectiveness of full-text, anchor text and mixture runs.

Table 6: Impact of page sampling on diversity of the TREC 2009 Diversity topics

		Qrels			Found in top 10			
		Sub-topics			Full-text		Anchor	
Percentage	Topics	Inf.	Nav.	Avg.	Inf.	Nav.	Inf.	Nav.
100.000	50	168	31	3.98	40	2	63	8
50.000	50	145	20	3.30	42	2	47	7
25.000	48	126	15	2.94	53	2	32	3
12.500	45	103	9	2.49	44	2	22	2
6.250	41	78	5	2.02	42	1	14	1
3.125	37	62	4	1.78	38	2	11	2

at the impact of inter- and intra-server links and found that inter-server links find more relevant pages for navigational topics than intra-server links, while the reverse is true for informational topics. Inter-server links tend to point to entry pages of sites, while intra-server links cover a much larger part of sites, which is in line with earlier studies [20]. In other words, for ad hoc search, intra-server links are at least as important as inter-server links.

On the right side of Figure 5 we see the impact of sampling pages on the effectiveness of anchor text and full-text. We only evaluate on the topics that have at least one relevant page in the sampled collections. Similar to the impact on the ad hoc performance of anchor text, the size of the collection plays a huge role. Performance drops as the collection becomes smaller, although the drop from 100% to 50% of the collection is less severe than for the ad hoc task. Oddly enough, the full-text run gets better with smaller collections. How can these observations be explained?

The impact of page sampling on the relevance judgements for the Diversity task are shown in Table 6. The number of informational sub-topics with at least one relevant document drops from 168 to 62, while the number of navigational sub-topics drops from 31 to 4. At the smallest samples, the Diversity relevance judgements have almost been reduced to ad hoc judgements. With only 1.78 sub-topics per topic, there is not much to diversify and the incurred penalty for retrieving pages on the same sub-topic is small.

What is the impact of anchor text on informational and navigational topics? The final four columns show the number of sub-topics for which the full-text and anchor text runs find relevant pages in the top 10. The diversity of the full-text run is hardly affected by the sample size, explaining why at smaller samples with less sub-topics available, its score goes up. The *Anchor* run finds both more informational and navigational topics in the full collection, showing it actually does better on the informational part of the topics than the *Text* run. However, it suffers greatly from the reduced collection size, especially for the informational sub-topics. Collection size plays a large role in the effectiveness of anchor text for informational search.

6. CONCLUSIONS

The history of scientific benchmarking for Web IR is plagued with the apparent contradiction between the experiences of Internet search engines, and the results of experiments at TREC [15, 16, 19, 33]. This led to Google's Larry Page calling the entire formal evaluation process "irrelevant" during a heated panel debate at the 2000 Infornotics Search Engine Meeting [31]. After several years of disappointing results at TREC Web Tracks, it was surmised that Web structure is simply not effective for ad hoc search tasks. TREC moved on to Web-centric tasks, where link topology, anchor text, and URL structure were proven very effective for navigational search, such as site finding and home page finding.

The availability of a new test collection, ClueWeb09 [7], which is a much closer approximation of the index of Internet Search Engines than earlier collections, prompts us to revisit the standing question of the importance of anchor text for ad hoc search. 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. More specifically, the pure anchor text runs lead to substantially higher precision than the full-text runs but the full-text runs have better recall. The straightforward combination of document and anchor text runs leads to significantly better scores throughout.

A negative finding is that link evidence like in-degree does not contribute to retrieval effectiveness. In contrast with the anchor text, link degree is ignorant of the query at hand, and its main use is in separating the authoritative or important pages from the less popular ones. An underlying difference with earlier collections is that the ClueWeb crawl is based on a PageRank/OPIC policy rather than the standard breadth-first strategy [12]. As a result, all pages in the collection have a relatively high level of 'importance,' and on top of that there is no additional value in link degrees.

The main focus of this paper is what makes the anchor text representation effective. Previous research pointed at the need of high (inter-server) link density [e.g., 4]. Our finding is that link density has little impact on anchor text effectiveness. Anchor text proved remarkably robust, and even with a small number of links it is effective for high early precision. Contrary to expectations, we find that intra-server link anchors are at least as effective as inter-server link anchors, even at equal density. Within a collection of high quality pages, such as the first tier of Web search engine indexes, the qualitative difference between inter- and intra-server links is minimal. But the greater quantity of intra-server links makes them more effective than inter-server links. Another factor is crawl or collection size, and ClueWeb09 is substantially larger than earlier testbeds. Our finding is that collection size has a big impact on the anchor text representations, affecting quantity, quality and effectiveness. A larger collection has more anchor-text covering a larger part of the collection. Especially within a crawl of high quality pages, more

links mean more high quality anchor text, leading to higher early precision than full-text search.

We also looked at the diversity task, which is a more Web-centric search task and does away with the notion that information is relevant no matter how often the user has seen it. The diversity task used the same topics and collection, but different judgments and measures. Our finding is that anchor text significantly outperforms full-text search, with greater differences and significance than for the ad hoc search task. This result also broadly confirms our findings for the ad hoc task. Anchor text is effective even at low link density; however, on smaller collections, the anchor-text covers an increasingly small part of the collection and loses its power. 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. Negative results for ad hoc informational search using Web structure have tainted the reputation of reproducible IR evaluation. The positive results in this paper may help to set the record straight. 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 Organization for Scientific Research (NWO, grants # 612.066.513, 639.-072.601, and 640.001.501).

We will make the link anchors and samples available upon request. For details see: `http://staff.science.uva.nl/~kamps/museum/anchors`.

REFERENCES

[1] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *WWW*, pages 280–290, 2003.

[2] B. Amento, L. Terveen, and W. Hill. Does 'authority' mean quality? predicting expert quality ratings of web documents. In *SIGIR*, pages 296–303. ACM, 2000.

[3] R. A. Baeza-Yates, C. Castillo, M. Marín, and A. Rodríguez. Crawling a country: better strategies than breadth-first for web page ordering. In *WWW*, pages 864–872. ACM, 2005.

[4] P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Inf. Process. Manage.*, 39(6):853–871, 2003.

[5] J. Callan, C. Yoo, and L. Zhao. Web08-PR Dataset, 2008. Project planning document.

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

[7] CMU-LTI. The ClueWeb09 Dataset, 2009. URL `http://boston.lti.cs.cmu.edu/Data/clueweb09/`.

[8] N. Craswell, P. Bailey, and D. Hawking. Is it fair to evaluate Web systems using TREC ad hoc methods? In *ACM SIGIR Workshop on Evaluation of Web Document Retrieval*, 1999.

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

[10] N. Craswell, D. Fetterly, M. Najork, S. Robertson, and E. Yilmaz. Microsoft Research at TREC 2009. In *TREC*, 2009.

[11] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *SIGIR '03*, pages 459–460. ACM, 2003.

[12] D. Fetterly, N. Craswell, and V. Vinay. The impact of crawl policy on web search effectiveness. In *SIGIR*, pages 580–587. ACM, 2009.

[13] D. Fetterly, N. Craswell, and V. Vinay. Measuring the search effectiveness of a breadth-first crawl. In *ECIR*, pages 388–399, 2009.

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

[15] D. Hawking. Overview of the TREC-9 Web Track. In *TREC*, 2000.

[16] D. Hawking and N. Craswell. Overview of the TREC-2001 Web Track. In *Proceedings of TREC-2001*, 2001.

[17] 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.

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

[19] D. Hawking, E. M. Voorhees, N. Craswell, and P. Bailey. Overview of the trec-8 web track. In *TREC*, 1999.

[20] D. Hawking, F. Crimmins, N. Craswell, and T. Upstill. How valuable is external link evidence when searching enterprise webs? In *ADC*, pages 77–84, 2004.

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

[22] J. Kamps. Effective smoothing for a terabyte of text. In *TREC*, 2005.

[23] J. Kamps. Experiments with document and query representations for a terabyte of text. In *TREC*, 2006.

[24] J. Kamps. Web-centric language models. In O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, and W. Teiken, editors, *CIKM*, pages 307–308. ACM, 2005. ISBN 1-59593-140-6.

[25] R. Kaptein, M. Koolen, and J. Kamps. Result diversity and entity ranking experiments: Text, anchors, links, and wikipedia. *TREC*, 2009.

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

[27] D. Metzler, J. Novak, H. Cui, and S. Reddy. Building enriched document representations using aggregated anchor text. In *SIGIR '09*, pages 219–226. ACM, 2009.

[28] M. A. Najork, H. Zaragoza, and M. J. Taylor. HITS on the Web: How does it compare? In *SIGIR '07*, pages 471–478. ACM, 2007.

[29] P. Ogilvie and J. P. Callan. Combining document representations for known-item search. In *SIGIR*, pages 143–150, 2003.

[30] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[31] C. Sherman. 'old economy' info retrieval clashes with 'new economy' web upstarts at the fifth annual search engine conference. *Information Today Newsbreaks*, 2000. `http://web.archive.org/web/20001217211000/www.infotoday.com/newsbreaks/nb000424-2.htm`.

[32] A. Singhal and M. Kaszkiel. AT&T at TREC-9. In *TREC*, 2000.

[33] A. Singhal and M. Kaszkiel. A case study in web search using TREC algorithms. In *WWW10*, pages 708–716, 2001.

[34] TREC. Text-REtrieval Conference, 2009. `http://trec.nist.gov/`.

[35] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *CIKM '06*, pages 102–111, New York, NY, USA, 2006. ACM.