Effectiveness Beyond the First Crawl Tier

Rodrygo L. T. Santos
rodrygo@dcs.gla.ac.uk

Craig Macdonald
craigm@dcs.gla.ac.uk

Iadh Ounis
ounis@dcs.gla.ac.uk

School of Computing Science
University of Glasgow
G12 8QQ Glasgow, UK

ABSTRACT

Modern Web crawlers seek to visit quality documents first, and re-visit them more frequently than other documents. As a result, the first-tier crawl of a Web corpus is typically of higher quality compared to subsequent crawls. In this paper, we investigate the impact of first-tier documents on adhoc retrieval performance. In particular, we analyse the retrieval performance of runs submitted to the adhoc task of the TREC 2009 Web track in terms of how they rank first-tier documents and how these documents contribute to the performance of each run. Our results show that the performance of these runs is heavily dependent on their ability to rank first-tier documents. Moreover, we show that, different from leading Web search engines, their attempt to go beyond the first tier almost always results in decreased performance. Finally, we show that selectively removing spam from different tiers can be a direction for fully exploiting documents beyond the first tier.

Categories & Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Performance, Experimentation

Keywords: Crawl Tiers, Effectiveness

1. INTRODUCTION

As the bandwidth, storage, and processing resources of any search engine to crawl and index the Web are limited, crawlers are guided by policies which focus on pages that are more likely to be relevant to user queries [6]. For instance, while breadth-first crawling finds high-value pages early [6], the OPIC (online page importance computation) measure is more often used in creating prioritisation policies [1].

The ClueWeb09 Web corpus[1] of 1.2 billion documents has provided the research community with a large sample of the Web, crawled in a language-specific manner. Figure 1 shows a schematic view of this corpus. For the purposes of the Text REtrieval Conference (TREC), the 500-million-document English subset of ClueWeb09 was designated 'category A'. Of this subset, a smaller subset with 50 million documents was designated 'category B'. The category B subset is reported to reflect the highest crawl priority, including a large number of high quality seed documents, as well as a snapshot of Wikipedia. As such, this subset roughly represents the first tier of a commercial search engine index.

[1] http://boston.lti.cs.gmu.edu/clueweb09/

CIKM'11, October 24–28, 2011, Glasgow, UK.
ACM X-XXXXX-XX-X/XX/XX.

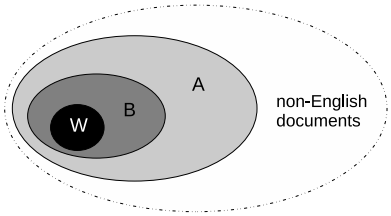


Figure 1: Schematic view of ClueWeb09, highlighting categories A and B, as well as Wikipedia (W).

For the adhoc task of the TREC 2009 and 2010 Web tracks, participants were encouraged to submit runs using both the category A and category B subsets of ClueWeb09. In this paper, we examine the effectiveness of the TREC 2009 category A adhoc runs,[2] particularly in light of the presence of first-tier documents from the category B subset. Our analysis reveals that B documents contribute substantially to the early performance of the A runs, while the impact of Wikipedia documents is less pronounced. On the other hand, going beyond the first tier can prove extremely rewarding, provided that spam is handled adequately.

In the remainder of this paper, Section 2 analyses the relevance assessments for the TREC 2009 Web track. Section 3 analyses the bias towards first-tier documents in the TREC submitted runs and two commercial search engines. Section 4 investigates the impact of the different tiers on the retrieval performance of these runs, as well as the impact of spam on each tier. Our conclusions follow in Section 5.

2. RELEVANCE ASSESSMENTS ANALYSIS

The adhoc task of the TREC 2009 Web track accepted both category A and B runs, which were pooled separately [3]. In this work, we focus on the category A runs and relevance assessments. In particular, these runs directly permit an assessment of the impact of first-tier documents (category B) on the effectiveness of rankings produced out of a larger crawl (category A). Table 1 reports the statistics of the category A, category B, and Wikipedia (denoted 'W') subsets of ClueWeb09, as well as the relevance assessments for the submitted category A adhoc runs to the TREC 2009 Web track. Besides absolute figures, we also show percent figures

[2] The runs submitted by the TREC 2010 participating groups were not publicly available at the time of writing.

denoting the fraction of B documents in A, as well as of W documents in B. Assuming that the order of the crawl was not correlated with relevance, we would expect 10% of the relevant documents to come from category B, as it represents 10% of the size of the category A subset. Likewise, we would expect 11.9% of relevant B documents to come from W. However, from the statistics in Table 1, there is a clear bias towards category B documents, as well as towards Wikipedia documents, both in terms of the documents assessed for relevance, and those judged relevant.

	number of documents		
set	crawled	judged	relevant
A	503,903,810	18,666	5,684
B ⊂ A	50,220,423 (10.0%)	8,183 (43.8%)	2,828 (49.8%)
W ⊂ B	5,957,529 (11.9%)	1,755 (21.4%)	669 (23.7%)

Table 1: Statistics of the relevance assessments for the TREC 2009 Web track category A adhoc runs.

3. RUNS ANALYSIS

To investigate the effect of first-tier documents on retrieval effectiveness, we firstly analyse their impact on the submitted A runs of the participating groups in the adhoc task of the TREC 2009 Web track. Since these runs were pooled to depth 12, we measure retrieval effectiveness at rank 10, which is also suitable in a Web search context [5]. Additionally, to enable the appropriate evaluation of unofficial TREC runs (i.e., runs that did not contribute to the TREC assessment pooling), all effectiveness measures are reported in terms of estimated precision at rank 10 (estP@10), as calculated using the Minimal Test Collections (MTC) method [2]. Besides the officially submitted category A TREC runs, we also analyse the performance of runs produced by two leading Web search engines (WSEs). In particular, for each of the TREC 2009 Web track queries, we obtain up to 1000 results from these WSEs using their public APIs.[3] The URLs retrieved by each WSE are then normalised and matched against those in the ClueWeb09 corpus. While this procedure is necessary to enable the reuse of the TREC Web track relevance assessments, the obtained rankings should be seen as a 'lower bound' of what these WSEs could have produced if they constrained themselves to the ClueWeb09 corpus. For each of the 37 submitted category A runs and the two WSEs runs, we measure the fraction of category B documents among the top 10, which we denote B@10. Figure 2(a) shows a scatter plot of estP@10 vs. B@10. Analogously, we measure the fraction of Wikipedia documents in the top 10, denoted W@10. The scatter plot estP@10 vs. W@10 is shown in Figure 2(b).

From Figure 2(a), we observe a wide spread of B@10 values, with a concentration around 0.1. This ratio is akin to the expected number of category B documents, which amounts to one tenth of all category A in ClueWeb09. However, this is markedly low in contrast to the aforementioned ratio of relevant documents from category B (i.e., 49.8% from Table 1). Moreover, we note that both WSEs (denoted with a • in the figure) exhibit B@10 around 0.5, which is much closer to the relevance expectation of 49.8%. Furthermore, from Figure 2(a), it appears that estP@10 is correlated with B@10. In particular, many of the runs with poor retrieval performance lowly rank category B documents. Conversely, other runs with higher B@10 are more likely to have higher performance. Indeed, the Spearman's ρ correlation coefficient between B@10 and estP@10 is 0.76, which attests the strength of this observation.

From Figure 2(b), we note that W@10 is concentrated around two separate regions: the first one (around 0.05) comprises runs that did not seem to deploy any special treatment for Wikipedia documents, while the second one (from 0.8 onwards) includes runs that apparently specifically targeted Wikipedia (indeed, one of the runs has W@10 equal to 1). Compared to B@10, W@10 shows a lower correlation with respect to the runs' attained estP@10 (Spearman's $\rho = 0.49$), which suggests that the impact of Wikipedia documents on the retrieval performance is less pronounced than that of first-tier documents in general (i.e., the B subset).

To illustrate these observations, Figures 3(a) and (b) show a breakdown of the top 10 retrieved documents and the relevant documents among these, respectively, averaged across all TREC 2009 Web track queries for each run. In both figures, the officially submitted runs and the two WSEs runs are ordered by estP@10, from worst to best, with the WSEs being the overall best. From Figure 3(a), we first observe that more effective runs indeed favour category B documents, with some of these runs also highly favouring Wikipedia documents. Nonetheless, the best runs (i.e., those by the considered WSEs) appear to deploy a retrieval strategy that better resembles the distribution of relevant documents in Table 1, with a balance between A and B documents. However, from Figure 3(b), we note that, of all retrieved documents, those from Wikipedia are the most likely to be relevant. Other category B documents (i.e., B-W) are the next 'safest' documents, while only the best performing runs are able to identify a substantial amount of relevant documents from the remaining A (i.e., A-B) subset.

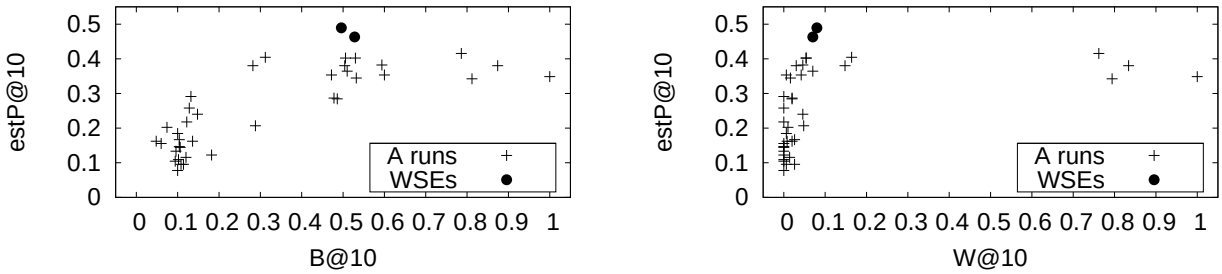
In addition to the relevance bias discussed in Section 2, this observation leads to an interesting finding: while being potentially very rewarding (nearly half of all relevant documents are not in category B), going beyond the first-tier of ClueWeb09 is also more challenging. In the next section, we investigate whether doing so is necessary for an effective performance, and what one could do towards this direction.

4. BEYOND THE FIRST TIER

Given the high prevalence of first-tier category B documents in the best performing A adhoc runs, we now assess how the presence of *non*-first-tier documents affected these runs. In particular, by contrasting these runs' ability to rank A and B documents, we can evaluate how well they perform beyond the first tier. To this end, Figure 4(a) shows a scatter plot contrasting the estP@10 of the A runs to the estP@10 of the same runs once all documents not in category B have been removed. Points above the $y = x$ line represent runs that are improved by the removal of the non-first-tier documents, while points under the line represent otherwise.

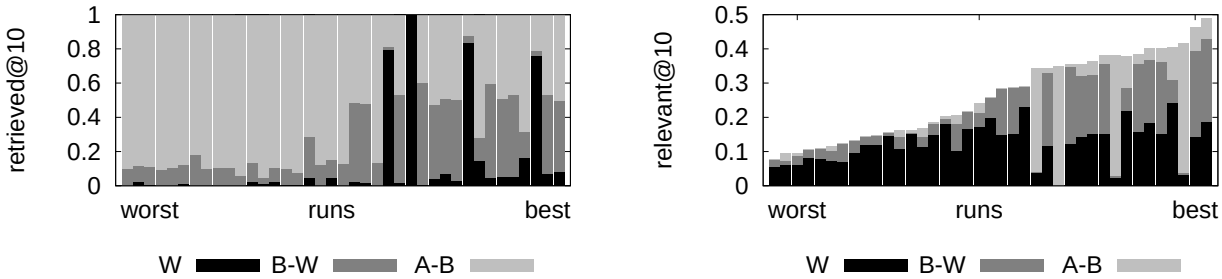
From Figure 4(a), we can see that almost all of the submitted A runs are improved by removing non-first-tier documents (i.e., documents in category A, but not in category B). Indeed, some runs with around 0.15 estP@10 are increased to 0.50, and only four runs are unaffected by the removal of non-first-tier documents. In contrast, we note that the runs by commercial search engines (denoted with

[3]Requests were sent anonymously to the US version of the WSEs, so as to isolate any customisation or localisation effects. All results were retrieved on February 7th, 2011.



(a) Impact of category B documents (b) Impact of Wikipedia documents

Figure 2: Retrieval performance for various amounts of (a) category B and (b) Wikipedia documents for the 37 TREC 2009 A adhoc runs (+) and two commercial search engines (•).



(a) Composition of documents retrieved (b) Composition of relevants retrieved

Figure 3: Composition of the top 10 documents (a) retrieved and (b) relevant across all runs. Runs are ordered by estP@10, from worst to best. The two best runs (the rightmost) are from Web search engines.

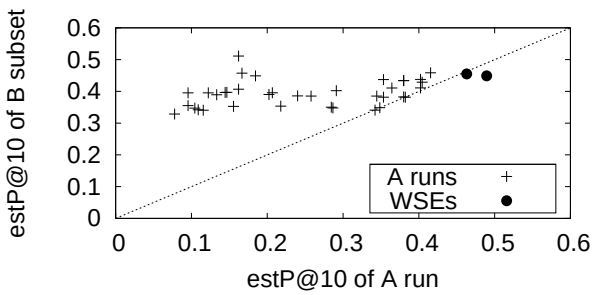
•) are both under the $y = x$ line, and as such, have been degraded by the removal of the non-first-tier documents. It seems that, in contrast to the TREC submitted runs, the commercial search engines are able to identify and competently rank relevant documents beyond the first tier.

Figure 4(b) shows the results of a similar experiment. This time, however, we analyse the importance of Wikipedia documents within the first tier. To this end, Figure 4(b) contrasts the subset of B documents in each of the considered runs to their subset of Wikipedia documents. Analogously to Figure 4(a), points above the $y = x$ line denote runs that benefit from ranking only Wikipedia documents, while points below the line denote runs that perform better when ranking all B documents. From Figure 4(b), roughly half of the runs fall on each side of the $y = x$. As a noticeable distinction, the markedly degraded performance of the two considered Web search engines when restricted to only ranking Wikipedia documents is due to their low coverage of these documents for the TREC 2009 Web track queries.

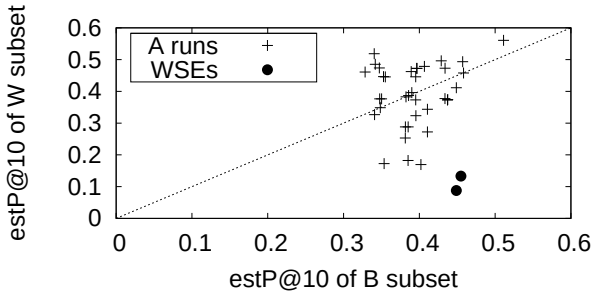
The results in Figure 4(a) show that the submitted TREC runs performed generally poorly at ranking non-first-tier documents. Analogously, from Figure 4(b), it is apparent that some runs cannot even effectively handle first-tier documents outwith Wikipedia. Cormack et al. [4] have shown that ClueWeb09 is severely affected by spam documents. To verify whether the results in Figures 4(a) and (b) are a mere reflection of the considered runs' inability to effectively handle spam, we repeat the same experiments after spam is filtered out. To this end, following Cormack et al. [4], we remove the 70% 'spammiest' documents from ClueWeb09 using their provided 'fusion' scores. The results of this investigation are shown in Figures 5(a) and (b).

From Figure 5(a), we observe that, although removing spam reduces the harm of going beyond the first tier, most runs still lie above the $y = x$ line, and are hence improved by restricting themselves to B documents. From Figure 5(b), we observe a different scenario, with most runs now lying under the $y = x$ line, meaning that going beyond Wikipedia and reaching out to the entire B subset is safer after spam removal. The different behaviours observed in Figures 5(a) and (b) suggest that spam removal, although overall beneficial, has a different impact on the different tiers.

To analyse whether this is the case, Figure 6 shows the impact of different spam removal configurations (averaged across all runs analysed in this paper) on the different strata of ClueWeb09: Wikipedia (W), non-Wikipedia category B (i.e., B-W), and non-category B (i.e., A-B). From the figure, as expected, spam removal can only hurt when applied to the W stratum, as Wikipedia is a spam-free corpus. When the B-W stratum is considered, only marginal improvements are observed for individual runs, with a decreased performance on average as spam is removed. Finally, substantial improvements are observed when the A-B stratum is considered. This result complements the findings of Cormack et al. [4], by showing that ClueWeb09 A indeed benefits from spam removal, but mostly from outside its first tier. As a result, a natural direction for improving retrieval performance on this collection is to selectively remove spam on the different strata (e.g., with a lenient spam removal threshold on the first tier, and a more aggressive threshold on the rest).

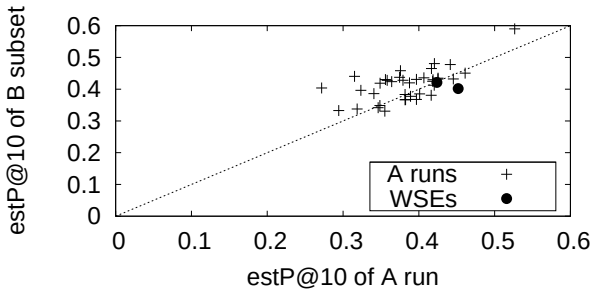


(a) Performance of category B documents

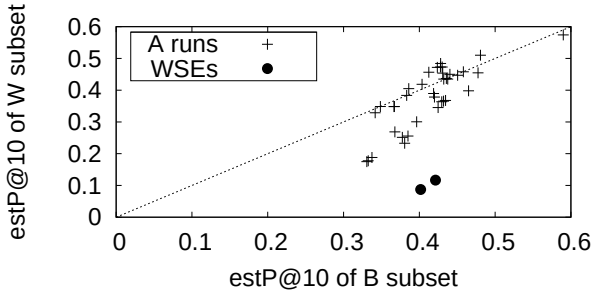


(b) Performance of Wikipedia documents

Figure 4: Retrieval performance of (a) category B vs. category A documents, as well as of (b) Wikipedia vs. category B documents, for the TREC runs (+) and Web search engines (•).



(a) Performance of category B documents



(b) Performance of Wikipedia documents

Figure 5: Retrieval performance of (a) category B vs. category A documents, as well as of (b) Wikipedia vs. category B documents, for the TREC runs (+) and Web search engines (•) after spam removal (70%).

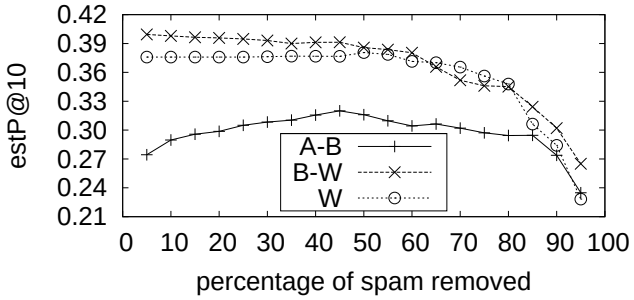


Figure 6: Per-subset average retrieval performance across all runs analysed in this paper as documents automatically classified as spam are removed.

5. CONCLUSIONS AND DISCUSSION

While it is well known that crawling order brings high-value documents early, the extent to which these documents are influential in the retrieval process requires further study. In this paper, we analysed the impact of the first-tier of the ClueWeb09 collection on the effectiveness of the submitted A runs to the TREC 2009 Web track adhoc task. Our experiments found that the official TREC runs that retrieved more first-tier documents were more likely to have higher effectiveness ($\rho = 0.76$). Moreover, by taking the extreme case of removing all non first-tier documents from these runs, we found that the effectiveness of almost all TREC runs was markedly enhanced. In contrast, when evaluating commercial Web search engines for the same queries, the removal of non first-tier documents was detrimental to effectiveness, suggesting that these commercial search engines are better at identifying relevant documents beyond the first crawl tier than the TREC systems. Finally, we have shown that the difference between the various tiers cannot be fully explained by the higher frequency of spam on the lower tiers. As a result, a possible direction for attaining an effective retrieval performance beyond the first tier is to apply different spam removal strategies on different tiers.

6. REFERENCES

[1] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez. Crawling a country: better strategies than breadth-first for web page ordering. In *Proc. of WWW*, pages 864–872, 2005.

[2] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *Proc. of SIGIR*, pages 268–275.

[3] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *Proc. of TREC*, 2009.

[4] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large Web datasets. *Inf. Retr.*, 2011.

[5] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: a study of user queries on the Web. *SIGIR Forum*, 32(1):5–17, 1998.

[6] M. Najork and J. L. Wiener. Breadth-first crawling yields high-quality pages. In *Proc. of WWW*, pages 114–118, 2001.