Measuring the Search Effectiveness of a Breadth-First Crawl

Dennis Fetterly[1], Nick Craswell[2], and Vishwa Vinay[2]

[1] Microsoft Research Silicon Valley, Mountain View, CA, USA
fetterly@microsoft.com
[2] Microsoft Research Cambridge, Cambridge, UK
{nickcr,vvinay}@microsoft.com

Abstract. Previous scalability experiments found that early precision improves as collection size increases. However, that was under the assumption that a collection's documents are all sampled with uniform probability from the same population. We contrast this to a large breadth-first web crawl, an important scenario in real-world Web search, where the early documents have quite different characteristics from the later documents. Having observed that NDCG@100 (measured over a set of reference queries) begins to plateau in the initial stages of the crawl, we investigate a number of possible reasons for this behaviour. These include the web-pages themselves, the metric used to measure retrieval effectiveness as well as the set of relevance judgements used.

1 Introduction

The Web is a very large collection of pages and search engines serve as the primary discovery mechanism to the content. To be able to provide the search functionality, search engines use crawlers that automatically follow links to web pages and extract the content over which indexes are built.

Crawling is usually described as a process that begins with a set of seeds, gathering new pages based on a pre-defined link exploration policy. When the crawler visits a page for the first time, it extracts all out-links on this page and adds them to the list of candidate links yet to be visited. At any given point, there are therefore two lists (a) all pages that have been visited (b) the 'frontier' consisting of pages the crawler knows of but has not yet visited.

If an exhaustive crawl was possible, the crawler would continue its operation until the frontier is empty. Given the size of the web, there are constraints that impose the need for the crawler to stop downloading new pages at a pre-defined point (for example, a limit on the number of pages in the index). It is therefore important to ensure that *good* pages get visited early on in the process. Past work have differed in terms of how they interpret the phrase 'good page'.

For example, [6] and [13] use link-based popularity metrics (like PageRank) to reflect the importance of a page. It is a reasonable expectation that in the presence of the early stopping criterion, greedily following links into popular URLs will lead to a good collection of pages. The RankMass of a crawler [7]

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 388–399, 2009.

formalises this notion by defining an index quality metric that is the sum of the PageRanks of its constituent pages.

A wide range of link exploration policies are available in the literature, addressing different motivations and subject to their respective constraints. Chakrabarti *et al* [4] consider "focussed crawling", the task of putting together a collection of topically related pages. The authors of [1] suggest limiting the depth to which websites are crawled to five. This conclusion, reached empirically from user session data, allows a crawler to obtain an even coverage across websites and domains on the rapidly expanding Web. The IRLBot Web-crawler [15] suggests domain-specific budgets for the number of pages crawled. Restrictions of this sort, which could be dependant on the domain's reputation, size, etc., ensure the scalability and efficiency of the crawler. Other criterion that have been considered when defining crawl selection methods are for example user-specific interests [18], the avoidance of spam [11] and wanting to obtain fresh versions of frequently changing pages ([5], [9]).

Breadth-first crawling, wherein pages are crawled in the order they are discovered, has been well-studied due to its relative simplicity. It has also been shown to yield high PageRank pages in the initial stages of the crawl [16]. We test this crawl policy on a larger scale than previous studies, and focus on its relationship with retrieval effectiveness. Our motivation for crawling the web-pages is to be able to service a search engine. By definition, a good crawl ordering policy is one that is able to stop potentially relevant search results from being crowded out by useless and redundant pages. In this paper, we consider monitoring the trajectory traced by retrieval effectiveness over the progress of the crawl.

Previous studies on the relationship between collection size and retrieval effectiveness [12] found that early precision improves as collection size increases. Hawking and Robertson's approach was to take a collection of 100 gigabytes containing 18 million documents, and measure early precision for corpus sizes of 100, 10 and 1 gigabytes. They found that the best early precision was achieved with the largest collection. It is noted that the crawling scenario is different because pages encountered early in the crawl are fundamentally different from those reached later on. This is in principle what would be expected from an effective crawling strategy. Our task in this paper is to investigate a range of metrics that reflect the behaviour of a breadth-first crawl at different stages.

2 The Initial Experiment

For the experiments described here, which extends the work reported in [10], we crawled 696,168,028 URLs between October 25, 2007 and November 28, 2007. Our crawl was started using the URL of the homepage of the Open Directory Project as the single seed. The crawl expanded out in breadth-first order. We wish to measure the retrieval-based utility of the crawled pages at chosen instances during the crawl, referred to here as *checkpoints* (we had 29 inspection instants).

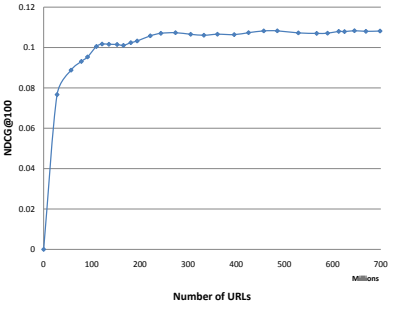

Fig. 1. NDCG @100

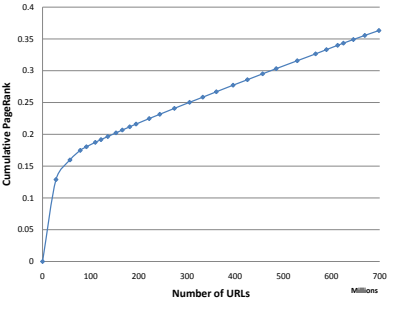

Fig. 2. Cumulative PageRank

A set of reference queries was constructed by sampling uniformly from the workload of the Live Search engine. These were matched with URLs judged on a 5 point scale for relevance: "Bad", "Fair", "Good", "Excellent" and "Perfect". Navigational results for a query (if any) were assigned the "Perfect" rating. We constructed a retrieval function that combines the well-known BM25 scoring method with an inlink prior using the method described in [8].

Our ranker was used to generate result sets of size 100 for each query in our reference set at each checkpoint. Using the relevance judgments available for each query, we calculated the search effectiveness achieved on a collection comprising of URLs crawled up until this checkpoint. The metric we used was Normalised Discounted Cumulative Gain (or NDCG) [14], which is a standard measure for web-based retrieval experiments and is used when graded (i.e., multi-level) relevance judgements are available. A *gain* is associated with each relevance category, a ranking algorithm is rewarded for not only retrieving documents with high gain but also for being able to place them high up in the ranked list. By tracing the value of NDCG through the checkpoints, we can estimate the utility of continuing the crawl. The results are provided in Figure 1.

The curve for NDCG has a spike at the start. Thereafter, the curve increases steadily, suggesting that the breadth-first crawl continues to reach pages that would improve user satisfaction for some time. Around the 225 million mark, this curve plateaus out indicating that there are diminishing returns, with respect to retrieval effectiveness. If we were to use NDCG as the primary decision making metric, according to Figure 1, the crawl should have been stopped at 225 million documents.

We believe that tracking the retrieval effectiveness through stages of the crawl is itself a novel experiment. However, the flattening of the NDCG curve so early in the crawl requires further investigation. Acknowledging that there might be multiple reasons for this behaviour, we next describe a series of experiments that look at alternate metrics to describe the state of the crawl at each checkpoint, hoping to tease out the underlying reasons for our initial observation. These are described in the next section.

3 Detailed Experiments

The plot of NDCG Vs crawl size is the result of interaction between many different factors. These include

- the link exploration policy
- the quality of the resulting corpus
- the judgements used for evaluation
- the ranker used for retrieval
- the metric used to represent effectiveness

We want to be able to measure the search-based utility of the corpus generated by a crawl, and this is effectively defined by the link exploration policy. If we are to determine how appropriate a breadth-first strategy is in being able to direct the crawl towards pages which will lead to high effectiveness, we need to systematically illustrate (and account for) the contribution of the remaining factors. We begin by considering the indexed URLs.

3.1 Link-Based Corpus Quality Metrics

As described earlier, we crawled just under 700 million pages starting from a specified seed. Once the crawl was completed, we constructed the link-graph of the entire collection of pages and calculated the PageRank [17] of each URL. We then calculated the cumulative sum of global PageRank values of all pages crawled up until each checkpoint. The word *global* is used to indicate that the PageRank was calculated on the final completed crawl (i.e., at the 29th checkpoint). Choosing this metric reflects previous use of link-based measures for evaluating crawl ordering policies. The resulting plot is shown in Figure 2. The cumulative PageRank has a steep rise at the start and this confirms previous work that a breadth-first search strategy obtains *good pages* at the very beginning [16].

On first look, we might think that Figure 2 contradicts the results in Figure 1. As opposed to the NDCG curve, the cumulative PageRank seems to be still on its way up when the crawl was stopped, indicating that there might have been some benefit in continuing the crawl. However, reaching a conclusion on corpus quality based on this metric is not so straightforward. All pages in the corpus are present because they have at least one incoming inlink, otherwise the crawler would not have reached this page. Also, the use of a uniform jump probability in the calculation of PageRank (we used a value of 0.15) means that every page in the crawl will have a non-zero PageRank, however small. Just by growing the crawl, we would expect the cumulative PageRank to increase.

The linear dependence between crawl size and the value of the metric is perhaps expected given published research, but is not in itself a positive vote for the breadth-first crawl policy. We might expect that this curve begins to flatten at some point, this saturation point can be guessed to be much larger than 700 million (the size of our crawl). Given that PageRank does not directly relate to

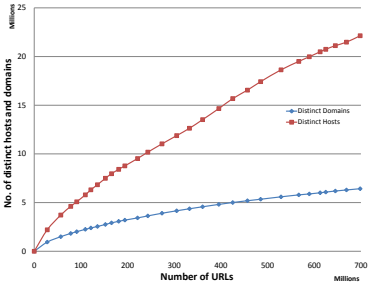

Fig. 3. Number of distinct hosts and domains with progression of the crawl

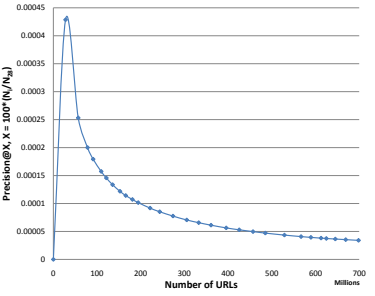

Fig. 4. Precision@X where $X = 100 * (N_i/N_{28})$ where N_i is the number of URLs in checkpoint i

search effectiveness, using it as a surrogate for corpus quality in this setting is perhaps not appropriate.

The Web is effectively infinite, and the information requests (represented by the queries a search engine receives) are also diverse. In order to be able to deal with queries of wide ranging topicality, we need a corpus that spans as broad a range as possible. We can achieve this by attempting to include in our index, pages whose content covers many topics. Alternatively, if we consider the website as being the atomic unit, one way of reaching a diverse set of pages is to ensure coverage over as many hosts/domains as possible. Such an objective can also be argued in terms of the search engine's *fairness* towards website owners. The difference between what constitutes a *host* and what is a *domain* is best illustrated by an example. We would consider "bbc.co.uk" as being a domain while "www.bbc.co.uk" and "blogs.bbc.co.uk" would be hosts.

In Figure 3, we plot the number of unique hosts and domains present at each checkpoint. We notice that while the number of hosts is increasing, the curve for number of domains is relatively flat. We posit that this might be symptomatic of using a breadth-first strategy because it is easy for a crawler to enter a domain with very many pages and *get stuck*. Even though the size of the crawl is increasing, the indexed pages might not be contributing towards the corpus quality, effectively representing an inefficient use of resources. Recently published research ([1], [15]) has drawn attention to this problem. Together with our experiments, this might indicate that a link-exploration policy that visits as many domains as possible might be necessary in order to achieve good search effectiveness.

3.2 Increasing Corpus Size

Hawking and Robertson considered the question of corpus size in detail in [12]. Their method was to take a large initial collection and sub-sample it to produce smaller collections. Experiments revealed that highest retrieval effectiveness (in particular, early precision) was achieved on the complete full corpus.

One of their hypotheses for explaining this observation was that there just weren't enough relevant documents in the smaller collections. Therefore, comparing Precision@X in a collection containing N_i documents to Precision@X in a larger collection with N_j documents is unfair when $N_i << N_j$. They suggested scaling the cutoff value X so as to account for the change in corpus size. Once this scaling has been performed, we would expect to see a flat Precision@X Vs Corpus Size plot. We repeated their experiment by considering Precision@100 on the complete full crawl, while the cutoffs for earlier checkpoints (i.e., X) were given by $X = 100 * (N_i/N_{29})$. Figure 4 provides the results.

As can be seen, the plot of Precision@X with scaled cutoff X falls with respect to increasing crawl size. We would not expect the conclusions of Hawking and Robertson to transfer completely to our scenario because crawling is inherently different from sub-sampling, but what does a fall in precision indicate? Relatively higher values of Precision@X in the early stages of the crawl indicate that the breadth-first link exploration strategy is achieving the desired objective of any crawling policy, that of identifying the desirable pages early on.

In terms of a measure of search-based corpus quality, the observed decreasing curve for precision is probably expected, the rate at which it falls is critical. Comparisons of breadth-first to alternate crawling methods will indicate how good each selection method is. The scaling of the cutoff value X (here linear with respect to corpus sizes) also needs to be revisited for the specific context of crawling.

3.3 Metrics for Retrieval Effectiveness

The retrieval experiment illustrated in Figure 1 was performed over a set of 7,248 queries sampled from the workload of the Live Search engine. Each of these queries had associated with them URLs whose relevance with respect to those queries had been obtained from human judges, there were 2,523,078 relevance judgements in total.

When the result sets for all queries were examined, we noticed that the number of documents from each label class saturates. We also observed that the fraction of URLs in result sets that are unjudged was sufficiently large to warrant the consideration of alternative retrieval effectiveness metrics. This is because NDCG by default assumes unjudged pages to be irrelevant, and this might explain the flattening of the curve. Even though NDCG has been shown to be stable with respect to incompleteness of judgements (e.g. [2]), we would like to ensure that we do not wrongly attribute saturating search effectiveness to a breadth-first crawl (Figure 1) purely because of our choice of metric.

To deal with the missing judgements, we considered infAP [19] and bpref [3]. Inferred AP, which is derived from the classical average precision measure, attempts to infer the relevance of unjudged documents in the result sets based on the pattern of relevance amongst judged URLs in the result rankings. Bpref is a metric based on the number of correct pairwise orderings in the returned rankings. Both these metrics require binary relevance judgements while our earlier NDCG experiments used graded judgements. We converted the available

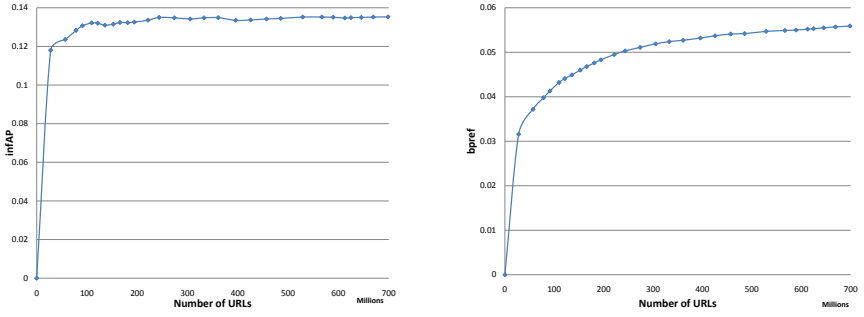

Fig. 5. Left: infAP - Relevance of unjudged URLs inferred from judged URLs in the result sets Right: bpref - URLs with missing judgements removed before evaluation

judgements into binary relevance labels by considering "Good", "Excellent" and "Perfect" as relevant, with all other labels being considered non-relevant.

We find that in Figure 5, infAP resembles the NDCG plot (Figure 1) very closely. While the flattening of effectiveness is not as pronounced for bpref as for the other metrics, there is evidence to indicate the diminishing utility of continuing the crawl.

The increasing bpref values in Figure 5 indicates that there are a few relevant documents that are crawled at the later checkpoints and our ranker brings them into the result sets. However, it is likely that these newly arriving relevant pages are at lower ranks of the result sets. This observation provides evidence to suggest that a measurement of retrieval effectiveness (as in Figures 1 & 5) is a convolution of two factors: (a) a crawl policy that is ineffecient in terms of reaching URLs that have been rated relevant (b) a retrieval function that does not rank the few relevant URLs that do make it into the corpus high enough to contribute to measured effectiveness.

We wish to point out that metrics designed to deal with incomplete judgements make a 'missing at random' assumption for the pages that are unjudged. A crawl policy that explicitly attempts to capture *good* pages early on in the crawl violates this assumption. That is to say, if we have an effective crawl policy, the ratio of number of relevant pages to the number of non-relevant pages is likely to be much larger in the beginning of the crawl before dropping off at the latter stages. This might indicate that for both metrics in Figure 5, retrieval effectiveness is being over-estimated in the early stages of the crawl.

3.4 Ranking Function

The retrieval effectiveness results described so far used a ranker that combined the BM25 score with an inlink prior. While a ranking function so produced might not reflect state-of-the-art, our assumption was that the trends of effectiveness results will be indicative of more sophisticated rankers. To ensure that the choice of ranking function does not influence the final conclusions, we conducted a further experiment.

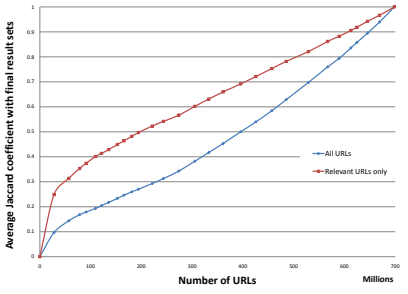

Fig. 6. Jaccard coefficient with result sets on complete crawl

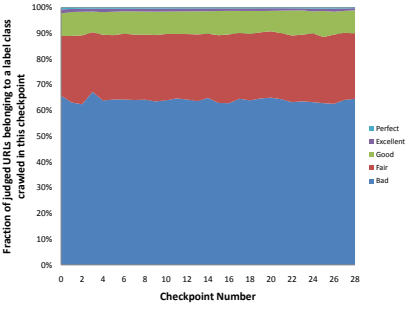

Fig. 7. Labelled documents picked up at different stages of the crawl

One possible reason for the flattening retrieval effectiveness is that the new incoming documents just do not make it into the result sets generated at that checkpoint. If they do not get ranked high enough in the results for the queries, they will not contribute to the NDCG, even if they are relevant. In order to check if this is the case, we calculated the average Jaccard similarity between result sets on the set of URLs in the current checkpoint and the final complete crawl. We repeat the calculation by considering only the relevant URLs in the result rankings, and compute the overlap between the relevant pages in top-100 of query results on intermediate checkpoints and the completed crawl. The plots of Jaccard coefficient Vs crawl size are provided in Figure 6.

We observe an almost perfect linear correlation between increasing crawl size and the average Jaccard coefficient between result sets produced on the final crawl and the intermediate checkpoints. This indicates that with the progressing crawl, we are continuing to add URLs to the corpus that tend to get ranked high (with respect to our reference query set) by the retrieval function. Contrasting the curve of Jaccard coefficient for all pages to that for relevant pages, we find that the breadth-first strategy identifies a large fraction of the relevant URLs in the final result sets very early on in the crawl. As with the results in Figure 4, this is a property that would be desirable in a crawling strategy.

In order to completely remove the effect of the ranking function used, we considered the entire corpus (rather than the result sets generated for each query) and counted the number of URLs belonging to each of our five judgement categories ("Bad", "Fair", "Good", "Excellent", "Perfect"). In Figure 7, we look at the set of judged URLs that were crawled between checkpoint i and $i-1$. A cumulative version of this plot, showing the number of URLs belonging to a particular label class at a particular corpus size, this is shown in Figure 8.

It can be seen from the figures that even though the number of "Bad" URLs increases at a rapid rate, there is an increasing trend for the presence of all other label classes as well. The crawling policy picks up relevant documents even in later stages of the crawl, and our rudimentary ranker places these in result sets (as verified by Figure 6). But the rate at which the relevant content is added

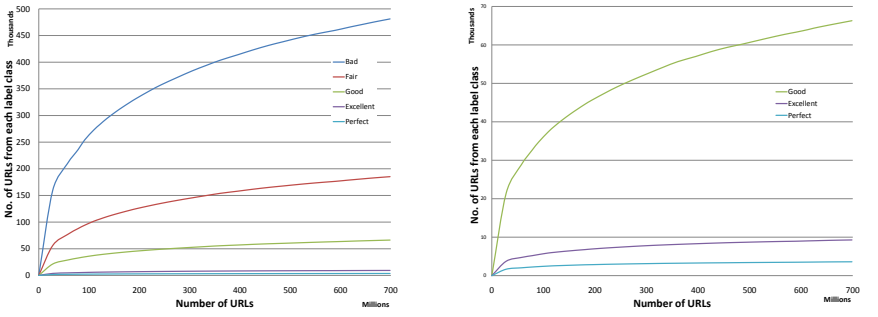

Fig. 8. Number of documents belonging to each label class in the corpus. On the right side is a zoom of the lower part of the Y-axis, to provide detail for the Excellent and Perfect categories.

to the corpus and the ranks at which they make it into result sets are not high enough to affect NDCG.

All the evaluations of the crawl that we have considered so far caused difficulty in segregating the performance of the ranking function from the quality of the crawl. In order to discount the contribution of choice of ranking function but still obtain a measure of search effectiveness, we choose to use "maxNDCG" as the metric to indicate the search-based utility of a corpus. This is based on the intuition that being included in the index is a pre-requisite for being returned as a result to a query, a *good* corpus selection method will make sure that the right documents have been chosen.

Given a corpus, in response to a query, the ideal ranking function would place all the highest gain pages at the top of the ranked list, followed successively by pages belonging to the classes associated with lower gains. The measurement of NDCG on this ideal ranking is therefore a function purely of the pages present in the corpus, which is in turn dependent on the crawl selection method used to generate the corpus. This leads us to maxNDCG for a corpus which rewards the crawl selection method for picking out the non-zero gain pages from the judgement set. We believe that due to the scale of our experiment ($> 7,300$ queries and $> 2,500,000$ judgements), maxNDCG will provide us with a reliable estimate of the quality of a given corpus, which is then used to characterise the crawl selection method that generated that collection of pages. The results are in Figure 9. It is worth noting that the relevance judgments were gathered on a different collection which is why the maximum possible maxNDCG line in Figure 9 does not have a value of 1.

Amongst the many metrics considered in this paper, we find that maxNDCG provides an indicator of search-related utility of a crawl that is easiest to interpret. Given our motivation to be able to evaluate the corpus generated by a particular link-exploration policy (breadth-first in this case), the challenge has been the design of a suitable experiment and use of a reliable metric. Assuming that NDCG provides some indication of user satisfaction, the calculation of

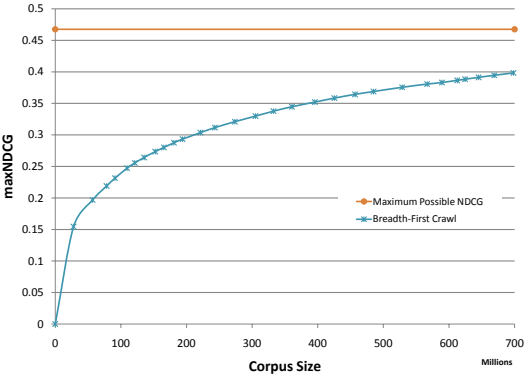

Fig. 9. maxNDCG for a breadth-first crawl

maxNDCG as described here is able to provide an indicator of potential future quality of a search engine that is serviced by this corpus.

If we are to judge the breadth-first strategy in terms of this metric, we find that it performs well, reaching over 80% of the maximum achievable NDCG (with respect to the reference set of relevance judgements), at a corpus size of roughly 700 million documents. Future work will compare this performance with alternate policies, hoping to identify those crawling methods that achieve high values of maxNDCG at lower corpus sizes.

4 Conclusions

In this paper, we considered the task of evaluating the corpus generated by a specified crawl ordering policy. This problem has been considered in the past, however the novelty of our experiment is that it looks at crawl corpus quality from the point of view of a search engine built on top of it. The particular method we evaluated was the simple breadth-first crawl, the advantage of using this method is that any prefix of the crawl is a valid breadth-first crawl itself. We were therefore able to ask questions about the benefits of continuing the crawl as a tradeoff between additional resources required for the pages versus potential increase in utility for end users.

To this end, we performed a large breadth-first crawl that successfully fetched ≈700 million URLs. Given that we wanted to measure the search-based utility of the crawl, we used a set of test queries and manually judged relevant pages as reference evaluation data. The rest of the paper described a series of metrics, and associated experiments, with an aim of factoring out as many experimental choices as possible, thereby obtaining a reliable measurement of crawl corpus quality.

Our starting point for the investigation was the use of a standard IR measure of retrieval effectiveness. A plot of NDCG versus crawl size showed diminishing

returns with increasing corpus size, with NDCG saturating after about 250 million pages. Observing the presence of a large number of unjudged URLs in our corpus, we measured infAP and bpref, retrieval effectiveness metrics designed to deal with missing judgements. As with NDCG, these methods also suggest that search effectiveness plateaus early in the crawl. We also computed the fraction of global PageRank at intermediate points in the crawl, confirming a previous result that a breadth-first crawl favours high-PageRank pages in its early stages. PageRank has been used in the past to evaluate crawling strategies, but does not directly relate to search effectiveness.

When examining the set of pages crawled, we found that there are a few relevant URLs that were being added to the corpus but our ranker failed to identify them. In trying to factor out the role of the particular ranking function used, we defined a measure that we call *maxNDCG* that is the effectiveness that an ideal ranker would achieve. The measure uses as input the collection of pages that comprise the corpus, and the set of relevance judgements against which retrieval effectiveness needs to be calculated. Since maxNDCG is purely a function of the corpus and is uninfluenced by experimental choices (other than the reference relevance assessments), we believe that this measure provides the most reliable indicator of search-based utility of a crawl strategy.

During the investigation of corpus quantity measures, we obtained some indicators of the behaviour of a breadth-first link exploration method. In favour of this crawl ordering strategy, it picked out relevant pages that tend to make it into result sets at a higher rate at the start of the crawl. Overall rates at which the *good URLs*, as defined by human relevance judges, were no higher or lower than other label categories at any stage of the crawl. Further evidence of the failings of the BFS crawl were provided by a sub-linear progress in the number of unique domains crawled. In terms of maxNDCG, we find that BFS gets close to the maximum possible, if this can be improved upon by other strategies remains to be seen.

Future work will compare alternate crawling methods towards the same motivation of trying to identify one that potentially leads to higher retrieval effectiveness. The design of crawl ordering strategies that are able to achieve high performance at low corpus sizes is an important problem. Tracing back from what users would perceive as positive characteristics of search results, all the way towards designing a crawl policy that ensures the inclusion of such pages into the corpus is therefore of great importance.

References

1. Baeza-Yates, R., Castillo, C.: Crawling the infinite web. Journal of Web Engineering 6(1), 49–72 (2007)
2. Bompada, T., Chang, C.-C., Chen, J., Kumar, R., Shenoy, R.: On the robustness of relevance measures with incomplete judgments. In: Proceedings of SIGIR 2007, pp. 359–366 (2007)
3. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proceedings of SIGIR 2004, pp. 25–32 (2004)

4. Chakrabarti, S., van den Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific web resource discovery. Computer Networks 31(11-16), 1623–1640 (1999)
5. Cho, J., Garcia-Molina, H.: The evolution of the web and implications for an incremental crawler. In: VLDB 2000: Proceedings of the 26th International Conference on Very Large Data Bases, pp. 200–209 (2000)
6. Cho, J., Garcia-Molina, H., Page, L.: Efficient crawling through URL ordering. Computer Networks and ISDN Systems 30(1-7), 161–172 (1998)
7. Cho, J., Schonfeld, U.: Rankmass crawler: a crawler with high personalized PageRank coverage guarantee. In: VLDB 2007: Proceedings of the 33rd international conference on Very large data bases, pp. 375–386 (2007)
8. Craswell, N., Robertson, S., Zaragoza, H., Taylor, M.: Relevance weighting for query independent evidence. In: Proceedings of SIGIR 2005, pp. 416–423 (2005)
9. Dasgupta, A., Ghosh, A., Kumar, R., Olston, C., Pandey, S., Tomkins, A.: The discoverability of the web. In: WWW 2007: Proceedings of the 16th international conference on World Wide Web, pp. 421–430. ACM, New York (2007)
10. Fetterly, D., Craswell, N., Vinay, V.: Search effectiveness with a breadth-first crawl. In: SIGIR 2008: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 755–756. ACM, New York (2008)
11. Gyongyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: VLDB 2004: Proceedings of the 30h International Conference on Very Large Data Bases, pp. 271–279 (2004)
12. Hawking, D., Robertson, S.: On collection size and retrieval effectiveness. Information Retrieval 6(1), 99–105 (2003)
13. Henzinger, M., Heydon, A., Mitzenmacher, M., Najork, M.: Measuring index quality using random walks on the Web. Comput. Networks 31(11), 1291–1303 (1999)
14. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (2002)
15. Lee, H.-T., Leonard, D., Wang, X., Loguinov, D.: IRLbot: scaling to 6 billion pages and beyond. In: Proceedings of WWW 2008, pp. 427–436 (2008)
16. Najork, M., Wiener, J.L.: Breadth-first crawling yields high-quality pages. In: WWW 2001: Proceedings of the 10th international conference on World Wide Web, pp. 114–118 (2001)
17. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
18. Pandey, S., Olston, C.: User-centric web crawling. In: WWW 2005: Proceedings of the 14th international conference on World Wide Web, pp. 401–411 (2005)
19. Yilmaz, E., Aslam, J.A.: Estimating average precision with incomplete and imperfect judgments. In: Proceedings of CIKM 2006, pp. 102–111 (2006)