Nullification test collections for web spam and SEO

Timothy Jones
Computer Science Dept.
The Australian National University
Canberra, Australia
tim.jones@cs.anu.edu.au

Ramesh Sankaranarayana
Computer Science Dept.
The Australian National University
Canberra, Australia
ramesh@cs.anu.edu.au

David Hawking
Funnelback Pty Ltd
Canberra Australia
David.Hawking@acm.org

Nick Craswell
Microsoft Research
Cambridge, UK
nickcr@microsoft.com

ABSTRACT

Research in the area of adversarial information retrieval has been facilitated by the availability of the UK-2006/UK-2007 collections, comprising crawl data, link graph, and spam labels. However, research into nullifying the negative effect of spam or excessive search engine optimisation (SEO) on the ranking of non-spam pages is not well supported by these resources. Nor is the study of cloaking techniques or of click spam. Finally, the domain-restricted nature of a `.uk` crawl means that only parts of link-farm icebergs may be visible in these crawls. We introduce the term *nullification* which we define as "preventing problem pages from negatively affecting search results". We show some important differences between properties of current `.uk`-restricted crawls and those previously reported for the Web as a whole. We identify a need for an adversarial IR collection which is not domain-restricted and which is supported by a set of appropriate query sets and (optimistically) user-behaviour data. The billion-page unrestricted crawl being conducted by CMU (web09-bst) and which will be used in the 2009 TREC Web Track is assessed as a possible basis for a new AIR test collection. We discuss the pros and cons of its scale, and the feasibility of adding resources such as query lists to enhance the utility of the collection for AIR research.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, selection process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

AIRWeb '09, April 21, 2009 Madrid, Spain.
Copyright 2009 ACM 978-1-60558-438-6 ...$5.00.

General Terms

Measurement

Keywords

Web spam, Test collections, Evaluation

1. INTRODUCTION

The rise in popularity of web search engines has caused an increase in the amount of web spam, aimed at manipulating rank in search engines. Web spam is a problem because it degrades ranking quality, and increases index size [19]. Web spam has received much interest recently, with two web spam specific test collections being created to test spam detection algorithms [10]. While these collections provide an excellent framework for testing spam detection algorithms, they do not provide easy evaluation of *nullification* algorithms (removing the effect of problem content such as web spam and excessive "search engine optimization" (SEO)).

To motivate the idea of nullification as opposed to removal, and to demonstrate that not all content that complicates ranking is also spam, we present the following example:

> An Australian football association published a website comprising several thousand genuine content pages. Each page was published using a template which included logos and links of its platinum and gold sponsors. Each of the gold sponsor links included anchor text along the lines of "National Australia Bank (NAB), proud sponsor of the Qantas Socceroos". In this real scenario, there are several thousand links targeting `nab.com.au` and including the word 'Qantas' in the anchor text. In response, a moderately sophisticated ranking algorithm may return `nab.com.au` (a bank) as a highly ranked answer to the query 'Qantas' (an airline).

This is clearly not an example of spam, since all sites involved are quite reputable and could not be eliminated from the crawl without harming results for other popular queries. However, achieving good search results requires the nullification of the effect of the thousands of template-driven links and their anchor text. Techniques for nullifying ranking degradation caused by obvious spam may also be effective against this kind of accidental template link spam, without the need to remove them from the index.

In addressing problem content on the web, there are four main problems to consider:

	Spam	Non-Spam
Removing pages	A	B
Nullification	C	D

A narrow view of adversarial information retrieval contains only problems A and C. However, if one is trying to remove spam and junk from a crawl, issues A+B are similar problems (eliminating bad pages) for which similar methods might be appropriate (link graph analysis, text-based classifiers etc). Similarly if one is interested in building a good ranking system then problems C+D share some similarities. These similarities are both in the type of problem being solved (ranking) and the kinds of solutions built (such as treating anchors differently, weighting link graphs etc).

As we see it, nullification of problem content is an essential component of successful ranking algorithms for the web. The goal of nullification is to optimise the quality of result sets presented to users, regardless of problem content present in the document collection.

2. RANKING AND WEB SPAM

While many readers will be familiar with web relevance ranking and the topic of web spam, we present a brief primer in the interest of completeness of argument.

2.1 Components of target algorithms

In order to investigate web spam, it is important to understand the components of ranking algorithms targeted. Partly because of malicious manipulation, commercial web search engines do not expose their ranking algorithms [5]. Web page ranking schemes almost invariably combine *static ranking functions* (query independent scores) and *dynamic ranking functions* (query dependent scores). Sorting documents by descending static score can allow early termination of query evaluation [23, 6], hence static scores are particularly important.

Link based schemes such as PageRank [25] are often considered the best way to achieve a good static ranking. However, combining other features such as user visit data, pure page content features, and general graph features (in-link count etc) can produce more effective results than PageRank alone [26]. In the following list, we briefly outline common features used in web search ranking.

- Content matching – matching a query to the content of a document. Examples include vector space, probabilistic and language models (See [3]). These approaches can be used to assign a score to a document reflecting the degree of relationship between the document's text and the query.

- Anchor text – Anchor text, meta-data, and other external textual annotations can also be scored for degree of match with the query [12, 27]. Unlike document content scores, scores derived from externally applied annotations reflect popularity and can be the subject of external spam, such as the once-popular “Google-bombing”[1].

- User behaviour – Web search engines accumulate vast quantities of search log data, recording sequences of query submissions and the URLs that were clicked on among the results presented. Click data can be used to build up a query-independent popularity profile for URLs in the collection [13]. Alternatively, queries resulting in clicks on a particular document may be associated with that document as a form of query-dependent evidence [31, 18]. Search engines with access to data from browser search toolbars may possess additional user interaction data, such as dwell times, printing actions etc., which can be used as a component of static scores.

- Link graph – PageRank [25], global HITS [8], and OPIC [1] are commonly used as static score components.

- Other features – Machine-learned functions such as those used by current major search engines may take into account hundreds of query-independent features such as URL length and structure, URL type, word density (homepage, privacy policy page etc.) and language [26].

- Regionalised features – Search indexes exposed to a particular region, e.g. Canada, may use feature vectors targeted at the particular region, in order to improve the quality of rankings for searchers in that region. For example, the query 'bank' submitted to Yahoo! UK is likely to return results for Barclays and RBS ahead of banks in other countries.

2.2 Web spam

In general, spamming may target any or all of the components of ranking algorithms. In this section, we briefly review the main categories of web spam techniques. An ideal collection for investigating web spam would contain examples of each type of technique. Gyöngyi and Garcia-Molina present three main types of web spam technique: *link spam*, *content spam* and *hiding techniques* [15]. We briefly describe each here. For a more detailed review, see their comprehensive paper.

- **Link spam** includes the practice of adding links to manipulate link based ranking schemes such as PageRank [25] or HITS [21]. Often link spam is organised into groups of heavily interconnected pages referred to as *link farms*. Link spam also includes manipulating anchor text, such as the misleading anchor text involved in “Google-bombing”.

- **Content spam** is altering the content of a document to include content aimed purely at the search engine. This includes techniques such as keyword stuffing and cloning good content from other areas of the web (such as wikipeida).

[1] e.g. By creating a large number of links, with appropriate anchor text, to a recent U.S. president's biography page, it was made to rank highly in response to the query, 'miserable failure'.

- **Hiding techniques** are designed to hide link and content spam from general web users. It can be as simple as white text on a white background, or more complex *cloaking* (serving different pages to search crawlers and users).

More recently, automated query traffic has also been seen. Buehrer *et al.* report that some of this seems aimed at engines who use click data as input to ranking, by repeatedly finding a specific target page in the search results and automatically "clicking" on it [7].

It is also worth mentioning search engine optimisation. SEO includes streamlining pages for search engines, such as ensuring anchor text and titles on pages are descriptive, and is generally encouraged by commercial search engines and appreciated by searchers. However, when optimisation becomes excessive it can impact content quality and user experience, becoming web spam.

Drawing the line between simple optimisation and web spam is difficult [10, 30], and there are several different definitions of the distinction [15, 14]. However, this line is clearer at extreme ends of the spectrum. Consider a site owner wanting to sell music who posts links to his site on music related forums, versus the same site owner creating hundreds of otherwise meaningless pages purely to point to his site. The first example may or may not be spam depending on a number of factors related to the forum communities and the relevance of his posts, whereas the second example is clearly spam. This difficulty of classifying optimisation versus spam was reflected in the labelling process for the UK-2006 collection. Disagreement among judges was high where at least one judge had classified a site as *borderline* (defined as heavily search optimised, but some useful content), whereas there was little disagreement where no borderline judgement had been made (all judges agreed a site was spam or non-spam) [10].

2.3 Detection and nullification

The majority of web spam literature to date has focused on the detection of web spam. However, some authors also discuss how to nullify the effect of detected spam, sometimes with a new ranking function [16, 22], or sometimes by adjusting the link graph around detected spam [30]. Because detection and handling need not be done in the same step, we make a distinction between *detection* (finding web spam pages) and *nullification* which we define as "preventing problem pages from negatively affecting search engine results".

It may seem intuitive that the most effective nullification is simply the removal of detected web spam, either from result pages or from the index. While simply filtering spam pages out of result lists can improve result quality [20], this simple approach does not remove all effects of spam. For example, link farms often target an otherwise non-spam page [15] which would not be demoted by this approach. Alternatively, removing all spam pages from the index would not correctly nullify spam links on otherwise good pages such as blogs [24], and may incorrectly punish sites that allow user submitted content such as forums and wikis.

3. EVALUATING SPAM NULLIFICATION

Here we review some of the ways spam nullification has been evaluated in the literature.

Wu and Davison present a graph based two stage spam detection algorithm, followed by a graph based nullification algorithm that produces a new ranking [30]. They note that as their algorithm does not consider page content, total removal of the detected pages is inappropriate. Their approach to nullification is to remove links between nodes labelled as spam. They then weight links on the cleaned graph to prevent mutual reinforcement, using a method proposed by Bharat and Henzinger [4]. Then, pages are ranked by the sum of weighted incoming links on the cleaned weighted web graph.

For evaluation, they use popular queries published by commercial search engines and also queries used by previous researchers. They select 20 of these queries and pool the top ten results from their approach, Kleinberg's HITS [21] and Bharat and Henzinger's BHITS [4]. For each query and pool, users were asked to rate documents on a five point relevance scale.

They also evaluate the effect of their graph modification on global static rank, using PageRank as an example. They use a 20M page collection and spam site blacklist obtained from a Swiss search engine. They then examine PageRank distribution of blacklisted sites before and after the nullification.

Gyöngyi *et al.* propose a version of personalised PageRank named TrustRank [16]. It replaces PageRank's uniform teleportation vector with one that links only to known good pages. In addition to evaluation by examining the distribution of known bad sites as ranked by PageRank and TrustRank, they report precision and recall values for known good documents after selecting n documents ranked by decreasing TrustRank scores.

Krishnan and Raj propose Anti-TrustRank, an inverted version of TrustRank that starts with a set of known bad pages, and propagates bad scores backwards along links instead of good scores forwards [22]. They report precision and recall of labelled spam pages for various ranks of Anti-Trust, but do not examine whether pages with low Anti-TrustRank scores are good.

Andersen *et al.* use insights from spam detection work to propose a spam resistant variation of PageRank called Robust PageRank [2]. They find the percentage of PageRank contribution to be a good feature for spam detection, so they modify PageRank to prevent incoming links from contributing above a certain threshold of PageRank to a target node. In this way, the spam detection phase is skipped, and spam nullification is performed immediately. They note that the ratio of Robust PageRank to pure PageRank is a good feature for spam detection. They evaluate their nullification in a similar manner to Wu and Davison, by examining the distribution of labelled nodes when ranked by PageRank and when ranked by Robust PageRank.

In a preliminary study [20], we built a two panel user experiment in which live queries were evaluated by real users, following the methodology of [29]. The interface allowed users to submit queries against the UK-2006 collection. For each query, users were presented with two panels, one pure results from the collection, and one with labelled spam filtered out at presentation time. We found a significant preference for the filtered result set, but did not continue with further experiments because many queries that users submitted appeared unanswerable within a domain limited web collection. Even UK specific queries had few answers in the collection (example "Jazz clubs in London"). This is likely to be because of the low number of hosts present in the UK-2006 collection.

3.1 Discussion of evaluation

Evaluating whether spam pages have been demoted to lower global ranks can be done with a spam labelled collection, using the popular method of examining distribution of spam pages before and after nullification. However, simply checking whether spam pages have lower global rank does not give information about the ultimate effect on live web search.

It is not clear that spam pages and relevant pages are at opposite ends of the relevance spectrum. In fact, some spam pages are created with relevant queries in mind. Therefore it is important to check that a spam nullification algorithm does not accidentally demote relevant pages. Additional evaluation could be performed using relevance labels paired with queries. If complete judgements are available, standard IR metrics such as MAP and NDCG can be used for evaluating relevance. These metrics can also be inverted to evaluate spam demotion (using the spam labels). In the case of large partially-labeled collections, metrics that are robust to incomplete judgements, such as infAP [32], would be more appropriate. The question of whether spam and irrelevant pages are identical is an interesting question for future research.

To make evaluation more convincing, user evaluation can be used alongside automatic evaluation. Both Wu and Davison's query pooling method [30] and our live search method [20] are appropriate. We would advise against our method building a live search index of a collection, in part because it is difficult to ensure that users' queries are answerable in the collection, but also because of the engineering difficulties in building a high quality index that responds in a reasonable time. The pooling approach has the advantage that searches do not need to be computed in real time. However, the pooling method does require the selection of queries for testing.

3.2 Properties of ideal collections

An ideal collection for examining web spam would have as much of the information available to spammers and commercial search engines as possible. This includes page content, the link graph, and some associated query and click data.

An ideal collection for detecting web spam would also have labelled spam and non spam pages.

In addition, an ideal collection for evaluating web spam nullification would also have some sample queries known to be affected by spam when state of the art ranking functions are used, and also some relevance judgements for those queries.

4. AVAILABLE WEB SPAM COLLECTIONS

There are two collections specifically tailored for testing web spam detection. They are the UK-2006 and UK-2007 collections, the most recent of which is used in the Web Spam Challenge track at the AIRWeb workshop. We describe their properties in this section.

4.1 UK-2006

The UK-2006 collection [10] contains roughly 80 million pages from roughly 11,000 hosts. The collection was created using a breadth-first crawl across hosts, with depth-first traversal within hosts. The crawl was restricted to links exclusively within the `.uk` domain, a crawl depth of 8 levels and no more than 50,000 pages from each host. The seed set was obtained from the Open Directory Project[2].

2,725 of the hosts in the collection have labels provided by 33 human judges. Each label is from the set {"normal", "borderline", "spam", "can not classify"}, using the guidelines[3] provided. Two automatic judges also contributed to the labelling. One marks controlled domains (such as gov.uk) as good, and the other marks pages in the Open Directory Project as good. These combine to give a total of 10,662 judgements, covering most of the hosts in the collection. However, most of these judgements have been provided automatically.

The availability of labels make it easy to evaluate web spam detection using the UK-2006 collection. However, it does not have any sample queries or relevance judgements. This makes it difficult to evaluate web spam nullification. It also does not have any associated click data. Furthermore, while the number of pages is quite high, the number of hosts is surprisingly low – only a tenth of the number of hosts in UK-2007. We assume that there must have been a technical or network problem during the UK-2006 crawl.

4.2 UK-2007

To address the issue of the low number of hosts in the UK-2006 collection, a similar collection was distributed with more hosts. The UK-2007 collection contains 105.8 million pages from roughly 115,000 hosts. Like the UK-2006 collection, some of the hosts also have human provided labels, using the same guidelines. There are 6,479 judgements in total, with 344 spam judgements, and 5,709 non spam judgements. There were no automated judges. Although there are far more human provided judgements in this collection, there are also far more hosts. Consequently only around 6% of the hosts have labels.

Similarly to UK-2006, the UK-2007 collection does not have click data, nor does it have sample queries or judgements for evaluation of spam nullification.

5. INVESTIGATING UK-2007

Since UK-2007 is the collection used in the current Web Spam Challenge, we use it in our investigation into the suitability of a domain restricted crawl for experiments in adversarial information retrieval.

5.1 Companion sites

Imposing a domain restriction on a crawl inevitably breaks up clusters of web domains operated by the same organisation or affiliated in some other way. To gauge the prevalence of this, we analysed the outgoing links from the UK-2007 collection and found more than **68,000** domains within `.co.uk` which were also represented in links to non-`.uk` domains. For example:

```
1click-insurance.co.uk -> 1click-insurance.com
1click2keys.co.uk -> 1click2keys.com
1click2keys-overseas.co.uk -> 1click2keys-overseas.com
```

[2] http://www.dmoz.org/

[3] http://www.yr-bcn.es/webspam/datasets/uk2006-info/

```
1click2keysoverseas.co.uk -> 1click2keysoverseas.com
...
3com.co.uk -> 3com.ch,3com.com,3com.cz,3com.de,
   3com.fr, 3com.nl,3com.se
...
abbott.co.uk -> abbott.com,abbott.de,abbott.dk,
   abbott.es, abbott.gr,abbott.ie,abbott.it,
   abbott.no
```

In our analysis, we used a library of structural heuristics to treat all the subdomains of a domain controlled by a single entity as a single domain (SECD, Single-Entity-Controlled-Domain). For example, `news.bbc.co.uk` and `www.bbc.co.uk` were considered sub-domains of `bbc.co.uk` while `www.layerone.com` was considered a sub-domain of `layerone.com`

The total number of non-`.uk` SECDs referenced in outgoing links (but not represented in the collection) was approximately 2.4 million.

We also observed very high link counts from some `.uk` domains to external domains not present in the collection. For example, we found at least 100,000 links from the left hand side of the following SECD pairs, to the right. In the domain-restricted collection, we obviously cannot see any links in the reverse direction.

```
layeroneuk.co.uk --> layeroneuk.com
lisburnontheweb.co.uk --> godaddy.com
theriddler.co.uk --> godaddy.com
homesandproperty.co.uk --> intelli-direct.com
iknow-cornwall.co.uk --> iknow-uk.com
iknow-yorkshire.co.uk --> iknow-uk.com
iknow-northwest.co.uk --> iknow-uk.com
iknow-lakedistrict.co.uk --> iknow-uk.com
bnn-online.co.uk --> uknws.com
...
```

The high number of hosts within the collection linking heavily to similarly named hosts outside the collection suggest that a domain restricted collection contains a very high number of *partial sites*, logically connected groups of pages for which only some of the pages are present in the collection. It is likely that a non-domain restricted crawl would not have this problem.

5.2 Answering popular queries

It is important that spam nullification is evaluated using queries targeted by spammers. We know these include popular queries [11]. If using the UK-2007 collection, it would make sense to use UK specific queries. In our earlier experiments, we anecdotally found many UK specific, user submitted queries unanswerable in the collection.

To evaluate whether UK specific popular queries are answerable using the UK-2007 collection, we obtained the top ten most popular queries for the UK from Google's 2008 Year End Zeitgeist[4]. We chose the 2008 list as it was UK specific; the 2007 list contained only globally popular queries.

We then submitted each query to two well known UK search engines, and obtained the top ten results from each, totalling 200 results. Table 1 shows the number of these results from hosts ending in `.uk`, and the number of the `.uk` results that are also present in the UK-2007 collection. Only 55 results (27.5%) were from hosts in `.uk`, and only 34 results (17%) were present in the UK-2007 collection. From this low number, and assuming commercial search returns good answers, it is likely that few good answers for popular UK specific queries are present in the UK-2007 collection.

	google.co.uk		uk.yahoo.com	
Query	in .uk	in UK-2007	in .uk	in UK-2007
facebook	0	0	0	0
bbc	6	5	8	5
youtube	0	0	0	0
ebay	6	2	5	2
games	0	0	2	1
news	8	7	6	5
hotmail	0	0	1	0
bebo	0	0	1	0
yahoo	0	0	0	0
jobs	7	4	5	3

Table 1: Number of results from .uk domains and Number of results present in the UK-2007 collection, for the first ten results from each popular query

If using user judgements over a country-code domain limited collection, the ideal test users would be located within the country. Since the most popular queries within the UK appear to be looking for pages outside of the `.uk` domain, it would be more appropriate to use a whole of web test collection, even if using only UK based users. As it is difficult for researchers outside the UK to recruit users from the UK, and as a whole of web test collection would suit live users anywhere in the world, a non domain restricted crawl would be generally more appropriate for evaluation involving popular queries.

6. POSSIBLE ALTERNATIVE COLLECTIONS

A number of web test collections have previously been created for research purposes and could potentially be candidates for use in adversarial IR research. Unfortunately, .GOV and .GOV2 collections most recently used in the TREC[5] Web Track [17] are not only restricted to a single domain (`.gov`), but to a domain which should be free of spam.

Earlier, the TREC Web Track used an 18.5 million general crawl carried out by the Internet Archive, known as VLC2, and small subsets known as WT10g and WT2g. Contemporaneous query logs from Excite and other search engines are available for these collections, but regrettably, the crawl is so old (1997) that the spam it contains cannot reflect present spam techniques. In 1997, PageRank and other link importance measures were not in use by search engines and therefore they were not targeted.

6.1 Extending the UK collections

It may be possible to extend the UK collections by crawling the rest of the partial sites present in the collection, and finding queries and relevance judgements that are both popular and contain answers in the extended collection. Unfortunately, this kind of extension would be difficult nearly two years after the most recent crawl, as the availability of many sites is likely to have changed.

[4] `http://www.google.com/intl/en/press/zeitgeist2008/#top`

[5] `trec.nist.gov`

6.2 Stanford WebBase

The Stanford WebBase Project[6] makes available many web crawls, some topic or domain focused and others unrestricted crawls of the general web.

General web crawls conducted monthly in 2008/2009 range from 61 million to 81 million pages. The most recent includes approximately 36,000 hosts. Each crawl is generated from the same site list, with no spam rejection. Full page content is available, with processed link information available on request.

Repeating the experiment in Section 5.2, 71 URLs (35%) are present in the most recent web crawl. This is considerably more than the 17% of the UK-2007 collection, even though the collection contains only a third of the number of hosts in the `.uk` collection. However, since the site list is constant each crawl, and spam sites are often short lived, the WebBase collections may not contain much spam. A cursory examination of the host lists for the two most recent crawls suggests this is the case.

6.3 Billion Page Crawl – web09-bst

Callan *et al.* [9] are currently producing a very large collection crawled in OPIC [1] order. To quote the web page, it is:

> A 25 terabyte dataset of about 1 billion web pages crawled in November, 2008. The crawl order was best-first search, using the OPIC metric. The crawl was started from about 25 million URLs that either i) had high OPIC values in a web graph produced from an earlier 200 million page crawl, or ii) were ranked highly by a commercial search engine for one of 16,000 sample queries in one of 10 languages. This dataset covers web content in English, Chinese, Spanish, Japanese, French, German, Arabic, Portuguese, Korean, and Italian.

The web09-bst collection deliberately contains multi-lingual content. This may make it interesting for spam research, as it is likely that web spam varies across languages and regions. For example, in China, cheaper domain name prices and relevance ranking methods different to those used for English text mean that the spam landscape is likely to be quite different. Spammers may also attempt to hide link farms by using language other than that of the target page, making human assisted detection by the search engine difficult.

The web09-bst crawl was not available at the time of writing, so we were unable to examine the possibility of answering popular queries using it, nor were we able to count the number of hosts present in the collection.

We summarise the collections we have discussed in Table 6.3.

6.4 Obtaining queries for evaluation

Research oriented toward measuring the adverse effect of spam and excessive SEO on search engine users cannot be conducted in the absence of sets of realistic queries and corresponding judgments. When selecting queries for evaluation of spam nullification, it is important to select queries of high interest to spammers. Chellapilla and Chickering [11] found the number of cloaked pages present in search results was higher when search queries were highly popular or monetisable (highly profitable advertisment terms). Since cloaking usually hides other spam techniques, it follows that popular and/or monetisable queries are common targets for spammers. It is important to note that by their very nature, popular queries are likely to have many high quality answers on the web. Because of this, it may be easier to fight spam that targets popular queries. However, the known higher proportion of spam pages targetting these queries means they are important for spam research.

Collection	Num pages	Num hosts	Spam labels?	Queries?
UK-2006	77M	11,000	Yes	No
UK-2007	105M	110,000	Yes	No
web09-bst	1B	Unknown	No	See Section 6.4
WebBase 04/06	92M	48,714	No	AOL
WebBase 01/09	75M	36,000	No	No

Table 2: Comparison of the suitability of various web collections for evaluation of spam nullification

Privacy concerns make search engine companies very reluctant to distribute query logs. Even though AOL removed IP addresses from the logs it distributed for research purposes in 2006, it soon became apparent that individuals were identifiable from sequences of queries they submitted [28].

The AOL query log consists of

> These records contain about 20M distinct queries submitted from about 650k users over three months (from March to May 2006). Each record is in the same format: AnonID, Query, QueryTime, ItemRank, ClickURL[7]

The AOL log is still accessible and could be used in conjunction with a contemporaneous non-domain-restricted crawl, such as the 92 million page crawl conducted by the Stanford WebBase in April 2006. However, some researchers are reluctant to work with this set of queries.

Commercial search engines do offer some query summary information, through services such as Google Year End Zeitgeist and Yahoo! Buzz[8]. Unfortunately these lists are very small, and tend to drop persistently popular terms in favour of more 'interesting' *mover and shaker* queries. These features mean they are not an ideal source for queries.

The new TREC Web Track plans to use the web09-bst crawl (see above) and also hopes to release Live.com query data in the form of the following histograms:

{query,count}	Raw query freq.
{query,url,count}	Freq. of click on URL after query
{query,query2,count}	Freq. of query pair in 10 min window

This proposed distribution avoids privacy concerns by discarding all session information and all data where frequencies are lower than a threshold. The data would be compatible with either web09-bst.v1 or a recent WebBase general crawl.

[6] `diglib.stanford.edu:8091/~testbed/doc2/WebBase/`

[7] Quoted from `tangra.si.umich.edu/clair/clair/qla.html`.

[8] `buzz.yahoo.com/`

6.5 Using large collections

Web collections comprising around 100 million pages or more pose major computational challenges to researchers with limited resources. We ourselves managed to index the UK-2006 and UK-2007 collections with around $US1.5k of hardware, and have achieved reasonable response times. However, a lot of engineering and indexing time was required. The task of re-indexing the collection in static-score order to produce better quality rankings is still before us.

Scaling by a further factor of ten to the scale of web09-bst will push the engineering required to a level beyond many researchers. To ameliorate the scale difficulties of the web09-bst collection, the web track organisers have indicated that a number of smaller derived files will be available. These may include ranked top 10,000 results for the test queries, which with the addition of other derived data (such as anchor text and link information) could be used to evaluate spam nullification without indexing the entire collection.

Alternatively, shared access to the collection and perhaps to pre-built indexes and data structures may be provided on large-scale computing infrastructure such as the Information Retrieval Facility[9], the Yahoo! M45 cluster[10] and the NSF/Google/IBM CluE program[11]. However, many researchers are likely to prefer to work on their own infrastructure.

In favour of a smaller collection:

- feasibility of complete spam labelling,
- fewer engineering difficulties for researchers in indexing, query processing, link graph calculations, and anchor text handling.

In favour of the largest available collection:

- link spam effects are likely to become much more visible in a crawl ten times as large, particularly with a best-first crawl.
- the billion page index may be big enough to contain most of the best answers to queries likely to be submitted in the course of user experiments.
- even given the latter point, a billion page crawl only represents a few percent of the crawls on which Google, Yahoo! and Live indexes are based.

7. DISCUSSION AND CONCLUSION

The UK-2006 and UK-2007 collections provide excellent frameworks for the evaluation of web spam detection. However, they do not well support the important task of evaluating web spam nullification.

We argue that a whole-of-web collection is more appropriate for evaluating spam nullification, and that such a collection needs to include a list of popular queries and relevance judgements for evaluation of spam nullification. For evaluation, we recommend the use one of or more of the techniques described in Section 3.1

We recommend the use of web09-bst, because its scale means that it is likely to contain much spam, and also more good answers for popular queries than existing collections. This will be easy to check using the method described in Section 5.2. For researchers unable or unwilling to index such a large collection, we recommend using a smaller subset of the web09-bst collection.

[9] `ir-facility.org`

[10] `research.yahoo.com/node/1884`

[11] `www.nsf.gov/pubs/2008/nsf08560/nsf08560.htm`

Acknowledgments

The authors would like to thank Jamie Callan for his regular correspondence about the web09-bst crawl.

8. REFERENCES

[1] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 280–290, New York, NY, USA, 2003. ACM.

[2] R. Andersen, C. Borgs, J. Chayes, J. Hopcroft, K. Jain, V. Mirrokni, and S. Teng. Robust pagerank and locally computable spam detection features. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 69–76, New York, NY, USA, 2008. ACM.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press/Addison-Wesley, New York, 1999.

[4] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 104–111, 1998.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[6] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *Proceedings of CIKM '03*, pages 426–434, New York, NY, USA, 2003. ACM Press.

[7] G. Buehrer, J. W. Stokes, and K. Chellapilla. A large-scale study of automated web search traffic. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 1–8, New York, NY, USA, 2008. ACM.

[8] P. Calado, B. A. Ribeiro-Neto, E. S. d. M. Nivio Ziviani and, and I. Silva. Local versus global link information in the web. *ACM Transactions on Information Systems (TOIS)*, 21(1):42–63, 2003.

[9] J. Callan, M. Hoy, C. Yoo, and L. Zhao. web08-bst.v1 web data collection, 2008. `http://boston.lti.cs.cmu.edu/callan/Data/#Web`.

[10] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. a. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.

[11] K. Chellapilla and D. M. Chickering. Improving cloaking detection using search query popularity and monetizability. In *Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, pages 17–24, Seattle, WA, August 2006.

[12] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In

Proceedings of ACM SIGIR 2001, pages 250–257, New Orleans, 2001. www.ted.cmis.csiro.au/nickc/pubs/sigir01.pdf.

[13] G. Culliss. User popularity ranked search engines, 1999. http://web.archive.org/web/20000302121422/http://www.infonortics.com/searchengines/boston1999/culliss/index.htm.

[14] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In *WebDB '04: Proceedings of the 7th International Workshop on the Web and Databases*, pages 1–6, New York, NY, USA, 2004. ACM Press.

[15] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[16] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proceedings of the 30th International VLDB Conference*, 2004.

[17] D. Hawking and N. Craswell. The very large collection and web tracks. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005. http://es.csiro.au/pubs/trecbook_for_website.pdf (ISBN 0262220733).

[18] D. Hawking, T. Rowlands, and M. Adcock. Improving rankings in small-scale web search using click-implied descriptions. In P. Bruza, A. Spink, and R. Wilkinson, editors, *Proceedings of ADCS 2006*, pages 17–24, Brisbane, December 2006.

[19] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[20] T. Jones, D. Hawking, and R. Sankaranarayana. A framework for measuring the impact of web spam. In *Proceedings of ADCS 2007*, December 2007.

[21] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[22] V. Krishnan and R. Raj. Web spam detection with anti-trust rank. In *Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, pages 37–40, 2006.

[23] X. Long and T. Suel. Optimized query execution in large search engines with global page ordering. In *Proceedings of VLDB 2003*, pages 129–140, 2003.

[24] G. Mishne. Blocking blog spam with language model disagreement. In *In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford, Santa Barbara, CA 93106, January 1998. dbpubs.stanford.edu:8090/pub/1999-66.

[26] M. Richardson, A. Prakash, and E. Brill. Beyond pagerank: machine learning for static ranking. In *Proceedings of WWW '06*, pages 707–715, New York, NY, USA, 2006. ACM.

[27] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *Proc. ACM CIKM'04*, pages 42–49, New York, NY, USA, 2004. ACM Press.

[28] X. Shen. Chronicle of aol search query log release incident, 2009. http://sifaka.cs.uiuc.edu/xshen/aol_querylog.html.

[29] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *Proc. CIKM*, pages 94–101, Arlington, Virginia, USA, Nov. 2006. ACM Press.

[30] B. Wu and B. D. Davison. Identifying link farm spam pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 820–829, New York, NY, USA, 2005. ACM.

[31] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing web search using web click-through data. In *Proceedings of ACM CIKM '04*, pages 118–126, 2004.

[32] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111, New York, NY, USA, 2006. ACM.