Improving rankings in small-scale web search using click-implied descriptions

David Hawking
ICT Centre
CSIRO
ACT 2601 Australia
david.hawking@csiro.au

Tom Rowlands
ICT Centre
CSIRO
ACT 2601 Australia
tom.rowlands@csiro.au

Matt Adcock
ICT Centre
CSIRO
ACT 2601 Australia
matt.adcock@csiro.au

Abstract *When a searcher submits a query Q and clicks on document R in the corresponding result set, we may plausibly interpret the click as a vote that Q is a description of R. We call the Q and R pairing a 'click description'. Click descriptions thus derived from search engine logs can be accumulated into surrogate documents and used to boost retrieval effectiveness in a similar fashion to anchor text.*

We investigate the usefulness of click description surrogate documents in processing queries for an external web site search service for four organisations. Using the mean reciprocal rank of best answers as the measure of performance, we show that, for popular queries, click description surrogates significantly outperform both anchor text surrogates and the original proprietary rankings. The amount of click data needed to achieve a high level of retrieval performance is surprisingly small for popular queries. Thanks to terms shared between queries, click description surrogates can answer queries for which no specific click data is available. We show a 92% improvement due to this effect for a set of lengthy, less popular queries.

We also discuss issues such as spam rejection, unpopular queries, and how to combine click description scores with other evidence. We argue the potential of click descriptions in non-web applications where link and anchor text evidence is unavailable.

Keywords Information Storage and Retrieval, Content Analysis and Indexing [Indexing methods]

1 Introduction

Many search engines not only log query submissions but also record details each time a user clicks on a search result. This 'click data' has previously been exploited in a variety of ways:

1. as low cost judgments in evaluating and tuning search engine performance [13, 14, 20, 1, 2]
2. as a query-independent page popularity score, used in similar fashion to PageRank [16] or indegree [5]
3. as a query-dependent popularity score [15]
4. to infer similarities between a pair of web pages on the basis that they were both clicked on in the same "session" [23, 24]
5. to infer descriptions of clicked-on web pages [24, 9] (When a searcher submits a query Q and clicks on document R in the corresponding result set, we infer that Q is a description of R.)

We focus on the potential of the last method, click-implied descriptions, treated in similar fashion to descriptions derived from anchor text, to contribute to effective search. An attraction of click-implied descriptions is that they may be used for collections in which there are no links and no anchor text.

The present study investigates the retrieval value of surrogate documents consisting only of concatenated descriptions inferred from clicks, where the inverse rank of the nominated best answer to each of a large set of queries is used as the effectiveness measure. Results for different types of query are presented for four different types of organisations; stock exchange, government, media and bank.[1]

2 Relation to past work

Document surrogates containing both anchor text and query associations have been found to improve retrieval effectiveness. Indeed, Web search engines have long made use of anchor text to improve result quality [4]. A variety of methods of using click data to improve ranking have been both described in the literature and exploited in commercial products.

2.1 Surrogates and supplements

For retrieval purposes, a text document may be *supplemented* with additional terms derived from external sources such as metadata, anchor text and so on. In the case of document *surrogates*, the additional terms form their own document which is used instead of the original. Retrieval may be based on scoring the surrogate collection or those scores may be combined with scores

Proceedings of the 11th Australasian Document Computing Symposium, Brisbane, Australia, December, 2006.

[1] Note that currently available test collections, TREC for example, are not distributed with applicable query logs and click data.

from the original collection. The following are examples of the use of surrogate or supplemented documents.

Sakai and Sparck-Jones [18] report that effectiveness in precision-oriented search is maintained when original documents are replaced by generic summaries during indexing. Craswell et al. [6] show dramatic improvements on homepage finding tasks of anchor text surrogates compared to the original documents. Scholer et al. [19] construct surrogates and supplements comprising controlled numbers of queries against which the original document ranked highly. They report increased accuracy in topic-finding searches but no benefit on a homepage finding task.

Amitay et al. [3] report equivocal benefit on the TREC-8 ad hoc task from supplementing documents using query reformulation sequences from a query log. The top k documents for the last query in a reformulation sequence are supplemented with the preceding queries in the sequence.

Hawking and Zobel [12] compare retrieval performance on a variety of query sets for surrogates comprising title-only, subject and description metadata and anchor text, in university and government site search. They use the rank of the best answer to a query as the criterion of performance. Overall, anchor text surrogates perform much better than the alternatives but the advantage is reported to be query set dependent.

2.2 Exploitation of click data

In the Chinese/English search engine WebGather, Lei et al. [15] linearly combine basic document scores with both link indegree and click frequency scores. Daily counts of distinct users who clicked are computed for (Q,P) pairs where Q is a query and P a web page. Furthermore, the count is multiplied by a factor designed to compensate for user reluctance to view more than the first page of search results.

The WebGather scheme compensates for changing user interests by combining the current day's scores with an attenuated aggregate of past days' scores. Finally, it also attempts to compensate for bias against recent documents. The authors enlisted ten users to 'mark' the system's performance and found an improvement over the baseline.

Smyth et al. [20] report the use of a similar hit-matrix in the context of a community-based metasearcher.

Joachims [13] presents a machine learning approach which adapts a search engine to a particular group of users. The author describes a method for training a retrieval function based around learning preference rules in the form 'for query Q, document D_a should be ranked above document D_b'. The author shows machine learning techniques are able to tailor a meta-search engine to a small group of users with similar interests.

Joachims et al. [14] discuss the reliability of the implicit feedback that can be derived from click data. The authors conclude that while click data is useful for relative relevance judgements it is problematic for absolute relevance judgements.

Click data was used in the past by the DirectHit search engine. Culliss (the DirectHit founder credited with the original idea) provided hints in [8] that DirectHit worked by 'monitoring' the sites users selected, boosting sites on which users dwell, penalising sites people don't select and rotating new sites in for review. Through a combination of these techniques, the system 'learns' from previous searchers.

Xue et al. [23, 24] study various ways of improving web search using an August 2003 click log for MSN Search.[2] This log includes data for approximately 63 million separate clicks.[3] The data covers 862 464 distinct queries.[4] They compare the methods and an Okapi BM25 baseline using a collection comprising only the webpages referenced in the click log. Xue et al. consider three methods in which Okapi scores are propagated to other pages based on 'co-visitation' relationships. The co-visitation similarity between two web pages is defined in terms of click frequencies:

$$CVS(d_i,d_j) = \frac{F(d_i,d_j)}{F(d_i)+F(d_j)-F(d_i,d_j)}$$

where $F(d_i)$ is the total number of clicks on d_j, regardless of query and $F(d_i,d_j)$ is the sum of clicks on d_i and d_j for queries associated with both documents.

Dmitriev et al. [9] use page 'annotations', both explicit and implicit, to improve intranet search results. They suggest that explicit annotations are expensive to produce, as they require users to produce them. On the other hand, implicit annotations, such as queries tied to pages that are clicked on in the result set for the query, while cheap to produce, are subject to users clicking on the wrong page. To mitigate these problems, they describe several other methods of extracting only the most valuable implicit annotations. Using the percentage of queries with a correct result in the top ten results as a measure, they show a significant improvement when using explicit annotations over the baseline, but no significant improvement with any of the implicit annotation schemes. The sample size used in the study, however, is quite small.

Agichtein et al. [1] explore the use of implicit feedback of many types, including click frequency and click rank. They compare the use of Okapi BM25 and neural network-based ranking methods, both with and without implicit feedback integrated as evidence and as a basis for reranking. They report significant gains with just click data and further gains with large vectors of implicit feedback. They observe that implicit feedback is particularly valuable for queries with poor original results. They do not address issues of spamming and the study is focused at all-of-web search.

[2] Then powered by Inktomi

[3] Personal communication

[4] After case-folding, stopping and stemming

In a related paper at the same conference, Agichtein et al. [2] go on to discuss dealing with 'noisy' user behaviour such as spam and clicks on irrelevant documents. They suggest that implicit features contain a background noise component which may be estimated by aggregating the behaviour of all users without regard to their query. Features where they suggest this may help include click frequency, dwell time and post-search behaviour such as clicks away from the original search page. They demonstrate this with a neural network-style system and report good results, with their best example delivering a recall of over 0.43 with a precision of over 0.67, which substantially outperformed their baseline.

Of most relevance to the present work, Xue et al. [24] compare the performance of click description surrogates ('NM - Naive Method') against two variants of the co-visitation method. They linearly combine surrogate scores with scores from the original documents. Using precision[5] and authority[6] measures for ten queries, they show that, when using all available click data, all three click-based methods perform roughly twice as well as the baseline. However, NM deteriorates more rapidly as the amount of click data is reduced.

2.3 Motivation for the present study

Most systems exploiting click data have been oriented toward whole-of-Web search, where click-spam is a potentially devastating problem. We wish to explore its applicability in small-scale enterprise contexts where spam is less of an issue and where document collections and click volumes are many orders of magnitude smaller.

The collection used in [24] is a very small subset, $\approx 5 \times 10^5$ pages, of the very large MSN Search collection, $\approx 3 \times 10^9$ pages in 2003 [21]. Consequently, link graph and anchor text information are incomplete and the effectiveness of the content-only baseline may not reflect values for the full collection. Also, the baseline is not the ranking against which the click data was generated.

In this paper, we attempt to show that even a naive method like that of [24] can be effective in webs of different scales. We evaluate with sizable query sets and we compare effectiveness relative to the rankings against which the clicks were generated. We also compare the relative value of anchor text and click-description surrogates in supporting effective retrieval for different classes of query. By using surrogates in isolation (following [6]) we hope to eliminate confounding variables. We also investigate the extent to which overlapping terms between click descriptions, query term overlap, helps or harms performance. Finally, we attempt to characterize the amount of click data required to achieve good performance.

[5]Precision at 20 documents retrieved

[6]The proportion of a pre-defined ten most authoritative pages which were returned in the top 20 results

3 Experimental method

In this study we use four crawled web corpora, each with two corresponding sets of queries. We compare five rankings: a baseline proprietary ranking and four rankings using simple Okapi BM25 scoring of surrogates: anchor text, two types of click descriptions, and document content only.

3.1 Datasets

Table 1 summarizes the document and click data used in our experiments. The data was crawled from externally facing websites. Anchor text for links pointing within the sites was available. The query and click logs were obtained from the production search facility for those sites. Clicks were recorded using a logging and redirection script. No attempt was made to hide the fact we were recording click data. In many enterprises, such a redirection script would not be required as the relevant information could be extracted from web server logs.

Our choice of corpora was constrained by the data available to us, but, fortuitously, the four organisations illustrate huge variations in crawl size and click density. Table 1 shows that the government corpus has one hundred times as much data but 430 times fewer clicks per page than the stock exchange corpus. The government collection includes hundreds of web hosts while the media collection includes fifteen and the stock exchange and bank include only one. During the time the logs were collected, approximately 23% of stock exchange pages received one or more clicks, but the comparable figure for the government collection was only 1%.

Queries submitted via advanced search interfaces were excluded to simplify analysis. Click entries in the log were lightly preprocessed to remove URL encodings ('+' and '%xx'). Operators were not removed, since they affect what is retrieved.

3.2 Test queries and judgments

Enterprise search systems are often judged (by searchers and purchasers) on the basis of their ability to rank the best answer to important queries at the top of the results list. As an example, consider the query 'Windows XP' submitted to the search facility on the Microsoft site. If the Microsoft Windows XP homepage doesn't appear at rank one, both users and site publishers consider this a search failure. Therefore, we evaluate by mean reciprocal rank of the best answer and use t-tests to check significance.

Table 2 summarizes the query sets used for evaluation purposes. For each corpus we obtained two sets of test queries and best answers.

Popular queries are the top queries ranked by frequency of querying, in some cases after excluding certain queries as explained in section 3.3.1. Judgements were made in collaboration with the relevant organisation where possible.

Sitemap queries are derived from the websites sitemap in a similar fashion to that described in

Table 1: Sizes of data sets and corresponding click logs

Collection	Pages	Clicks	Clicks/Page	Distinct pages clicked	Distinct queries clicked
stock exchange	2.2×10^4	1.9×10^5	8.6	5 038	23 427
government	2.3×10^6	4.6×10^4	0.02	22 055	17 539
media	7.6×10^5	9.3×10^4	0.13	44 647	30 379
bank	2.7×10^3	6.7×10^3	2.47	1 176	3 576

[10]—entries in the sitemap become test queries and the links become the corresponding best answer. This is a low cost evaluation method in which judgments are again made by the publishing organisation.

A peculiarity of the stock exchange site is that often three letter stock codes are used as queries. For example, seventy seven of the top one hundred queries are three letters long and the vast majority of these are codes. The best page for all of the codes is, according to the stock exchange, a CGI script with the code as a parameter. Such cases (perfectly answerable by a simple mapping) are not particularly interesting. Consequently, we tested the sixty seven most popular non-stock code queries. Click-through data with three letter queries is, however, included in the click surrogates.

3.3 Baseline rankings

The rankings against which click events were logged were generated using a proprietary retrieval system which is understood to make use of metadata, anchor text, and web measures such as link counts and URL properties. The production index was constantly updated over the time studied.

For our baseline condition, we indexed a crawl which was used in all experiments. As a result, the production rankings against which clicks were generated may differ from the baselines reported here. However, we are confident that large ranking perturbations would be relatively unusual. The baseline was generated by the proprietary software and used similar indexing and query processing parameters to the production service.

3.3.1 Complications with production baselines

Facilities provided by commercial search tools may interfere with good science in various ways. Two examples are as follows.

Table 2: The query sets used for evaluation purposes. Average lengths are given in words and include stopwords.

Collection	Test type	Queries	Ave. Length
SE	Popular	67	1.31
SE	Sitemap	491	2.98
Gov	Popular	87	1.25
Gov	Sitemap	430	4.32
Bank	Popular	49	1.63
Bank	Sitemap	256	2.73
Media	Popular	45	1.6
Media	Sitemap	35	2.31

The histogram of clicks shown in Figure 1 shows discontinuities at ranks which are multiples of ten for the stock exchange, corresponding to the default numbers of results per page. Similar discontinuities are present in the corresponding plot for the government service, but at multiples of twenty, as the result pages are longer. This apparent reluctance of users to click on the next page of results, and wait, may further reduce the chance of a low-ranked result being promoted through clicks.

Another complication is the availability of links triggered by a query but generated from a mechanism such as a look-up table maintained by the search administrator, rather than from the normal ranking mechanism. Examples of this mechanism include the targeted advertisements on major Web search engines and the "Editor's Choice" links on `search.microsoft.com`.

Regrettably, in our data, clicks on such results were not logged. If the nominated best page for one of our queries was the subject of such a mechanism, it would be unlikely to receive any click descriptions. Accordingly, we eliminated from analysis the queries which were subject to such short-cuts.

3.4 Creation and scoring of surrogates

In effect, surrogate documents are created by assembling the words into documents which take the place of the original ones, along the lines of [6]. All the surrogates of a particular type are indexed as a collection. Surrogate documents are then ranked using the familiar Okapi BM25 formula [17]. For the content-only surrogate, the settings from [17] $k_1 = 2.0$, $b = 0.75$ were used, as appropriate for normal text.

Hawking et al. [11] argue that length normalisation makes little sense with anchor text surrogates. The

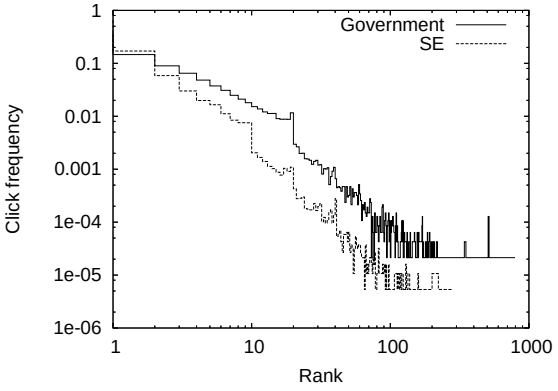

Figure 1: Reduction in click frequency as result rank increases

Table 3: Size of each collection in millions of bytes.

	Original	Anchors	Click words
bank	101.9	1.7	0.1
stock exchange	177.3	5.4	1.3
media	12 468.1	495.2	0.6
government	51 649.6	1107.5	0.7

same argument applies here as click data also provides a form of voting. Accordingly, length normalization was disabled by setting $k_1 = 2.0$, $b = 0.0$, for anchor text and click description surrogates.

Stemming and query expansion were not employed.

The sitemap pages from which the sitemap queries were derived give an obvious bias toward anchor surrogates in those cases. Consequently, the pages from which the sitemap tests were derived were removed from the index for those tests.

Four types of surrogates were studied:

Anchors —anchor text between `<a>` and `</a>` tags only from all incoming links, after following redirects.

Content only —original document, including title but excluding metadata, HTML markup, JavaScript, image tags and so on.

Click words —the surrogate for the document referenced in each pre-processed click log entry has each of the corresponding query words appended to it.

Click tokens —each distinct test case is represented by a single unique token (see Table 4). In the query log, each query that is identical to a test case is replaced with the equivalent token for the purposes of the surrogate document. This nullifies the effect of any query word overlap.

3.5 Surrogate collection sizes

As may be seen in Table 3, the sizes of the click description and anchor text surrogate collections tend to be very much smaller than the original. The difference is dramatic in the case of government where the click words surrogate corpus is around 0.001% of the size of the original.

3.6 Word overlap

Both in the case of anchor text and of click descriptions, it is possible that changes to rankings may arise from the sharing of words between different descriptions. For example, if clicks for the query 'oil corp' and clicks for the query 'stock price' hit the same document, then the word-based click description surrogate for that document contains a full match to the query 'oil corp stock price' even if no clicks have been recorded for that query. This may, or may not, be useful.

Table 4: Hypothetical example queries with example tokens

Token	Query
q1	abcd airlines
q2	oil corp
q3	wxyz corporation
q4	abcd

We investigate the effect of overlap in click descriptions by comparing effectiveness differences between the click words and click tokens surrogate collections listed above. Test queries submitted to the token collection are, of course, expressed as the appropriate tokens.

4 Experiments

In this section we describe the aims and conditions of each experiment and report results.

4.1 Experiment 1—Click effectiveness

The aims of this experiment are as follows:

- to investigate whether rankings based on click surrogates are capable of improving on the original baseline
- to compare the performance of Okapi BM25 rankings over content, anchor text and click surrogate collections
- to confirm whether patterns of results are the same on four corpora of very different sizes and on query sets devised in very different ways

Each set of queries are run against content-only, anchor text and click words corpora. The results are shown in Figures 2 and 3. We evaluate by mean reciprocal rank of best answer, and use t-tests to check significance. Several key observations may be made:

- For the bank, government and stock exchange corpora, the click desciptions ranking significantly outperforms the baseline for popular queries ($p \leq 0.01$)
- For all corpora, the click descriptions ranking is significantly outperformed by the baseline for the sitemap query sets. ($p < 0.01$)
- For the bank, government and stock exchange corpora, the click words surrogates are significantly more useful than anchors when processing popular queries. ($p < 0.01$)
- For all collections other than stock exchange, the click words surrogates are significantly more useful than content when processing popular queries. ($p < 0.01$)
- The apparent advantage to click words over content for popular queries, in the case of the stock exchange corpus, is not significant. ($p > 0.05$)

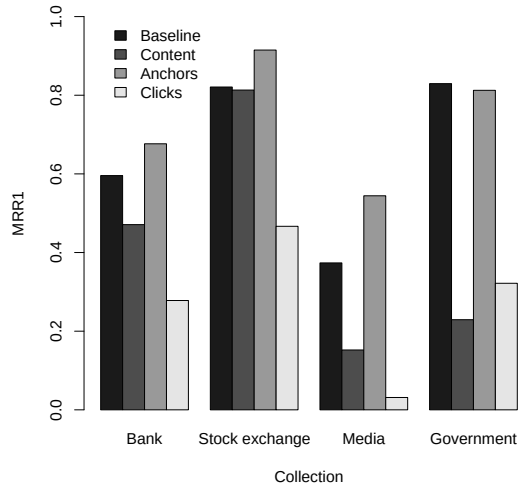

Figure 2: Sitemap tests: $p < 0.01$ except media content

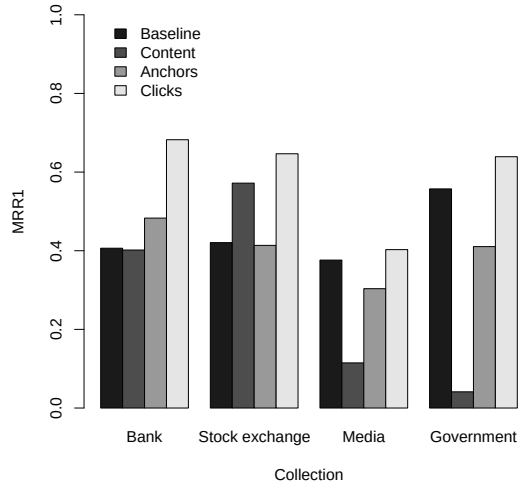

Figure 3: Popular tests: $p < 0.01$ except stock exchange content where $p < 0.09$ and media baseline and anchortext where $p > 0.1$

- In the case of the sitemap queries, click descriptions are significantly less useful than anchors on all four collections. ($p < 0.01$)

- Click descriptions in sitemap tests are significantly more useful than content only for government but significantly less useful in the other cases. ($p < 0.01$) There is no significant difference in the case of media ($p > 0.05$).

- There is a very large difference in performance for the content only surrogates, both in absolute terms and relative to the other surrogates. Performance is much higher on the smaller collections.

Table 5: Results from Experiment 2. The difference in the second row is significant. ($p < 0.01$)

Query Set	Words	Tokens
Popular	0.576	0.554
Sitemap	0.457	0.237

4.1.1 Discussion

It seems likely that the entries in a sitemap would tend to use the same language as the rest of the site, leading to higher performance for anchors and content only surrogates for sitemap based tests. On the other hand, familiarity with official nomenclature is likely to be imperfect among site visitors, tending to lead to queries (and clicks) for short queries, such as 'health', rather than to 'ministry of health and ageing'. Table 2 shows that the average length of sitemap queries is greater than popular queries; more than twice as long for stock exchange and government. The result of this may be that sitemap based queries are more specific, submitted less often and less capable of deriving benefit from click evidence.

4.2 Experiment 2—Query word overlap

The aim of this experiment is to determine the extent of harm or benefit due to word overlap across queries.

Two query sets are run against the click words and click tokens surrogate collections for the stock exchange corpus. Results are shown in Table 5. For the popular queries, there is no significant difference ($p > 0.05$, Wilcoxon signed rank significance test) between the scores, indicating that query word overlap is not important. By contrast, for the sitemap queries, query word overlap increases the MRR1 score by 92%, presumably because the exact sitemap queries are rarely submitted.

4.3 Experiment 3—Quantity of click data

The aim of this experiment was to investigate the relationship between performance on test sets and the amount of click data available. We did this by building click words corpora from randomly chosen samples of the click data.

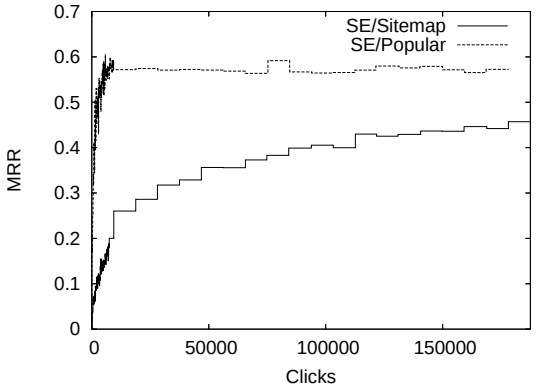

Figure 4: This figure shows the effect of varying the number of clicks used to assemble the surrogate document.

Each point in Figure 4 represents the average of two samples of approximately the same size. As may be seen, performance increases with sample size, approaching an asymptotic limit. As few as 4000 clicks are sufficient on the popular stock exchange queries to achieve an MRR score of 0.5.

The sitemap queries also start increasing rapidly but slow, approaching the final MRR less far less rapidly than the popular queries. It takes over 100 000 clicks before the sitemap queries show an MRR of 0.4 and they never reach 0.5.

5 General discussion

Our results show that click description surrogates achieve best results for popular queries. This is no doubt largely due to the larger amount of click data available for those queries, but it is a plausible supplementary hypothesis that the very short nature (average less than 1.5 words) of the popular queries makes it hard for other ranking schemes to reliably identify the best answer.

A major problem with ranking based on clicks is the potential for "fraud". By clicking repeatedly, a user can bias the ranking to favour a result. This method of artificially up-weighting results is believed by some to have lead to the demise of the DirectHit search engine [20], but we expect query dependent usage to be less susceptible than query independent popularity counts. An analogous ranking inflation technique for anchor text is "Googlebombing" [22].

Inside an enterprise, there would be no financial incentive to dishonestly manipulate rankings by clicking. Public facing enterprise websites similarly offer no incentive to manipulative clickers. Unfortunately, public websites such as a stock exchange or media site are likely to be among the exceptions to this rule.

Regardless of incentive, there may be techniques which can be used to counteract artificial clicks without excessively damaging result quality. Each click is recorded with a time, source IP address, referrer URL, query, destination and so on. It may be possible to use regularity of clicking and source address filtering with heuristics to filter out fraudulent clicks. Additionally, similarly to Web Gather [15], cookies could be used to limit the number of clicks recorded per user. Further, an asymptotic ranking function could take into account 'over-clicked' documents. The use of background click frequency, as discussed in [2] may also offer some immunity to click spam. This area of adversarial IR seems worthy of further study.

It may be possible to compensate for the trust bias discussed in [14]. Some initial experiments have been conducted by gradually downweighting clicks made on results ranked in the top ten. Initial results are equivocal. Further investigation is necessary.

Scholer et al. [19] introduced mechanisms to limit the size of query association surrogates in order to stay within defined storage limits. Here, the size of the click surrogates arising from the query log data are small relative to anchors surrogates and very tiny compared to the original data. Click surrogate sizes increase with increasing query volume, but size can easily be controlled by sampling or using a temporal sliding window on the logs.

Like link count, PageRank, anchor text, and other recommendation techniques, there is a potential bias in click data against new content. However, unlike simple query-independent click popularity counts, scores derived from click surrogates are query dependent and therefore generally capable of more rapidly responding to changes. Consider the hypothetical case of a highly controversial government report 'the Cierpinski Report', whose publication on a government website causes a massive increase in popularity of the query 'Cierpinski Report'. Provided that the baseline search algorithm was able to return the report at a point in the ranking visible to some searchers, click evidence linking the query with the document would build up. Using simple query-independent click frequency, it would take a very long time to compete with other popular documents. On the other hand, using query-dependent click data, the desired answer can potentially very quickly overtake other candidates. The temporal sliding window method helps to limit bias against new content in the case of queries with ongoing popularity.

Problems of link redirection and of links to documents eliminated by a duplicate detector are much less of a problem for click data than for link measures and anchor text. On the other hand, unlike anchor text, click evidence cannot provide descriptions of documents external to the crawl.

To achieve success for all types of query and for tasks other than best-document finding, we believe that click description scores should be combined with scores from anchor text and content and with query-independent measures. Xue et al. [24] report a simple fusion technique for combining with content while Craswell et al. [7] propose methods for combining query dependent and query independent evidence.

Click description surrogates depend upon an initial baseline ranking and are to some extent limited by its failings. We have found that click evidence is capable of promoting documents from deep in the original result rankings and also, due to query word overlap, to respond to queries never previously typed. It is not clear whether occasional random perturbations of rankings as practised by DirectHit would lead to better click descriptions.

Although we have demonstrated substantial improvements over the baseline (and over anchor text surrogates) for popular queries on a best-document finding task, the greatest potential gain from click data may lie in non-web environments, where link measures and anchor text are unavailable. Examples of such environments include library subject catalogues, and search of personal or corporate email and office documents.

6 Conclusion

We show that click-implied description surrogates alone can support good performance on best-document finding tasks in four very different webs. Using these surrogates, a mean reciprocal rank score of over 0.5 is achieved for popular queries in three out of four test corpora.

For the popular query sets, rankings based on click surrogates alone significantly outperform the original baseline ranking for three out of four corpora. They also outperform the ranking derived from Okapi BM25 scoring of anchor text surrogates.

We find that surprisingly little click data is necessary to achieve good results for popular queries and that performance on the best answer finding task approaches an asymptote once sufficient data is available. Click description surrogates are consequently very small, leading to efficient calculation of retrieval scores.

Comparison of query word and query token surrogates for the stock exchange sitemap set of queries shows a major benefit (92% relative gain in MRR) due to query word overlap. By contrast, there was no significant benefit on the popular query set, comprising much shorter queries with much more click evidence available.

Many interesting avenues await further research including: development of more sophisticated analytical models; methods for combining click surrogate scores with other ranking information; determining whether there is additional value in using click data as query independent evidence; spam rejection techniques; and investigating the use of clicks in non-web applications.

References

[1] Eugene Agichtein, Eric Brill and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proc. SIGIR*, 2006.

[2] Eugene Agichtein, Eric Brill, Susan Dumais and Robert Rango. Learning user interaction models for predicting web search result preferences. In *Proc. SIGIR*, 2006.

[3] Einat Amitay, Adam Darlow, David Konopnicki and Uri Weiss. Queries as anchors: selection by association. In *Proc. HYPERTEXT*, 2005.

[4] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. WWW*, 1998.

[5] J. Carriere and R. Kazman. Webquery: Searching and visualizing the web through connectivity. In *Proc. WWW*, 1997.

[6] Nick Craswell, David Hawking and Stephen Robertson. Effective site finding using link anchor information. In *Proc. SIGIR*, 2001.

[7] Nick Craswell, Stephen Robertson, Hugo Zaragoza and Michael Taylor. Relevance weighting for query independent evidence. In *Proc. SIGIR*, 2005.

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

[9] Pavel A. Dmitriev, Nadav Eiron, Marcus Fontoura and Eugene Shekita. Using annotations in enterprise search. In *WWW*, 2006.

[10] David Hawking, Nick Craswell, Francis Crimmins and Trystan Upstill. How valuable is external link evidence when searching enterprise webs? In *Proc. ADC*, 2004.

[11] David Hawking, Trystan Upstill and Nick Craswell. Towards better weighting of anchors (poster). In *Proc. SIGIR*, 2004.

[12] David Hawking and Justin Zobel. Does topic metadata help with web search? *JASIST*, 2006. (To appear.).

[13] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proc. ACM KDD*, 2002.

[14] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. ACM SIGIR '05*, 2005.

[15] Ming Lei, Jianyong Wang, Baojue Chen and Xiaoming Li. Improved relevance ranking in webgather. *Journal of Computer Science and Technology*, Volume 16, Number 5, 2001.

[16] Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998. `http://dbpubs.stanford.edu:8090/cgi-bin/makehtml.cgi?document=1999/66`.

[17] S. E. Robertson, S. Walker, M.M. Hancock-Beaulieu and M. Gatford. Okapi at TREC-3. In *Proc. TREC-3*, 1994. NIST spec. pub. 500-225.

[18] Tetsuya Sakai and Karen Sparck-Jones. Generic summaries for indexing in information retrieval. In *Proc. SIGIR*, 2001.

[19] Falk Scholer, Hugh E. Williams and Andrew Turpin. Query association surrogates for web search. *JASIST*, Volume 55, Number 7, 2004.

[20] Barry Smyth, Evelyn Balfe, Jill Freyne, Peter Briggs, Maurice Coyle and Oisin Boydell. Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction*, Volume 14, Number 5, 2005.

[21] Danny Sullivan. Search engine sizes, 2005. [Online; accessed 24 Jan 2006] `http://searchenginewatch.com/reports/article.php/2156481`.

[22] Wikipedia. Google bomb—Wikipedia, The Free Encyclopedia, 2006. [Online; accessed 21 Jan 2006; `http://en.wikipedia.org/w/index.php?title=Google_bomb&oldid=36057937`].

[23] Gui-Rong Xue, Shen Huang, Yong Yu, Hua-Jun Zeng, Zheng Chen and Wei-Ying Ma. Optimizing web search using spreading activation on the clickthrough data. In *Proc. WISE*, Volume LNCS 3306, 2004.

[24] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi and WeiGuo Fan. Optimizing web search using web click-through data. In *Proc. CIKM*, 2004.