Relative Effect of Spam and Irrelevant Documents on User Interaction with Search Engines

Timothy Jones
Australian National University
Canberra, Australia
tim.jones@anu.edu.au

David Hawking
Funnelback Pty Ltd & Australian National Univ.
Canberra, Australia
david.hawking@acm.org

Paul Thomas
CSIRO
Canberra, Australia
paul.thomas@csiro.au

Ramesh Sankaranarayana
Australian National University
Canberra, Australia
ramesh@cs.anu.edu.au

ABSTRACT

Meaningful evaluation of web search must take account of spam. Here we conduct a user experiment to investigate whether satisfaction with search engine result pages as a whole is harmed more by spam or by irrelevant documents.

On some measures, search result pages are differentially harmed by the insertion of spam and irrelevant documents. Additionally we find that when users are given two documents of equal utility, the one with the lower spam score will be preferred; a result page without any spam documents will be preferred to one with spam; and an irrelevant document high in a result list is surprisingly more damaging to user satisfaction than a spam document. We conclude that web ranking and evaluation should consider both utility (relevance) and "spamminess" of documents.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

General Terms

Measurement, Human Factors

Keywords

Web Search, Web Spam, Adversarial Information Retrieval

1. INTRODUCTION

Web spam documents—documents designed to achieve an unfairly high rank for some page or pages—are a problem in web search. The topic of web spam dectection has received much recent attention. However, relatively little attention has been paid to the problem of spam nullification—understanding how to correctly deal with spam documents once detected [3]. In order to understand how to appropriately nullify spam documents, it is important to investigate the relationship between spam and relevance. Understanding this relationship is also valuable when evaluating spam nullification. For example, if users see spam as equivalent to irrelevant documents, then the effectiveness of spam removal can be measured using traditional evaluation methods.

In this paper we examine the relationship between spam and relevance, to determine whether spam labels are required when measuring spam nullification. We present the first study we are aware of which compares user reaction to spam documents to user reaction on other low value documents (in this case irrelevant documents). We find that not only do users prefer spam results to irrelevant results, but surprisingly a click on a spam document can sometimes increase the user's satisfaction with a result set.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

2. PREVIOUS WORK

Users behave differently when faced with spam and non-spam documents. Spam documents receive nearly all of their visits from search engines, whereas non-spam documents are also visited via a variety of other means (direct links, bookmarks, etc.) [6]. Similarly, when a user views a document from a spam site, they are unlikely to navigate to many other documents within that site.

This user browsing behaviour can be exploited to detect spam. In [7] and [9], the user browsing graph is used to detect spam. This exploits the property that users are unlikely to navigate to spam pages from other pages in the web graph. A similar approach is to use the view and click graphs obtainable from query logs [2].

Here we also consider non-spam, irrelevant, documents and ask: do spam and irrelevant documents induce similar user behaviour, or do users react differently to the two?

3. EXPERIMENT DESIGN

Participants in our investigation were assigned information gathering tasks and asked to complete them using our search engine. Unknown to users, each task was completed using a different search engine: either a high quality engine, or one of six degraded engines. Task order was controlled in order to avoid carry-over effects.

Including a practice task, users completed eight tasks using seven different engines. After each task, users reported

their success in completing the task. Users were told they could end the task if they had been searching fruitlessly for several minutes.

3.1 Controlling for engine quality

In this experiment, live web search performance was degraded by a predetermined amount. To do this, we inserted low value documents into the results provided by the high quality search engine (Yahoo!'s "BOSS" API, `developer.yahoo.com/search/boss/`). Our model for estimating the degredation to NDCG@10 scores caused by various patterns of inserted documents is described in more detail in [8].

From this model, and assuming that both spam and irrelevant documents provide zero gain to the user, we test three expected degradations to NDCG@10 score: 0.2, 0.4 and 0.8. The pattern to achieve a degradation of 0.2 is insertion at ranks 2 and 3; that is, after inserting zero-value documents at ranks 2 and 3, we expect the NDCG@10 score to drop by 0.2. For 0.4 we insert at ranks 1, 3, 8, and 9, and for 0.8 we insert at ranks 1, 2, 4, 5, 6, 7, 9, and 10.

Insertions. Three irrelevant engines were constructed by inserting irrelevant documents into result pages from the high quality engine. We denote the least degraded engine *(i:2,3)*, indicating that *i*rrelevant documents were inserted at positions 2 and 3. The other engines were *(i:1,3,8,9)* and *(i:1,2,4,5,6,7,8,9,10)*.

In order to provide a comparison with the irrelevant documents, spam engines were also created. These engines were identical to the irrelevant engines, except spam documents were inserted instead. These engines were *(s:2,3)*, indicating that *s*pam documents were inserted at positions 2 and 3, *(s:1,3,8,9)* and *(s:1,2,4,5,6,7,8,9,10)*.

Sourcing spam and irrelevant documents. Selecting spam documents using a commercial search engine is difficult, as we can expect the commercial engine to be taking steps to remove spam from the results. However, the ClueWeb'09 collection is a recent collection that has spam labels provided [4]. We retrieve spam documents for insertion by submitting users' queries to a spam-only index of ClueWeb'09 Category B. Because we use their real queries, the spam documents will appear to be answers to a user's search. Documents selected in this manner appear to be low quality, clearly spam, and will have some overlap with the query.

Selecting irrelevant documents is more complex. Several methods were considered, but we found the best results were achieved by adding low meaning terms (such as buzzwords) to the query. We selected a joke buzzword generator available online (`www.1728.com/buzzword.htm`), and used the output to add terms to the user's query. This approach works surprisingly well when using a commercial search engine. Results appear targetted to the query, are non-spam but are of little use.

3.2 Logging

Logging was performed using JavaScript served along with the results page. Several features were logged, including query text, depth of result page, clicks on search results, scrolls on search pages, the index of visible results after scrolls and focus switches away and back to the search result page (which occur when the user switches tabs or uses the browser back and forward navigation buttons).

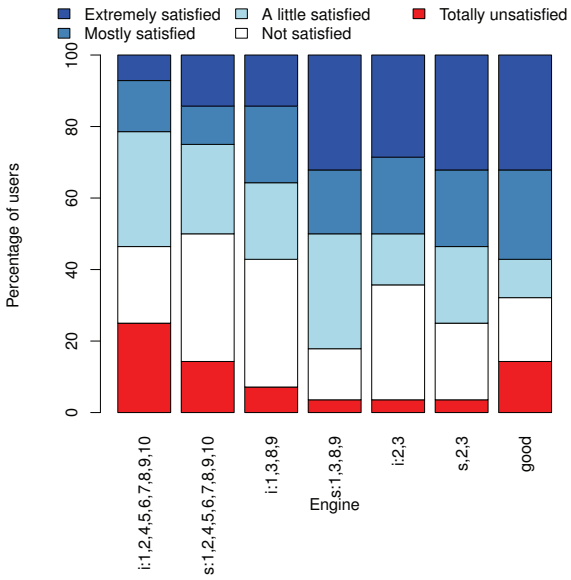


Figure 1: User reported satisfaction (lowest value used for each user).

As already discussed, upon task completion, users self-reported task success on a three point scale, which we report as *success*, *partial success* or *failure*.

We also obtained user judgements on the result set quality. To prevent these judgements from affecting user behaviour, explicit ratings were only solicited after all tasks were completed. After all tasks were completed, users were shown previous queries from each task, and asked to rate their satisfaction with the first result page on a five point scale. Up to five queries were randomly selected from each task, for a maximum total of 40 queries rated per user. This total was less if a user had submitted less than 5 queries for any task.

3.3 Tasks used

Participants were given eight information gathering tasks, each of which required several searches to complete. Three of the tasks were created for students in Canberra, three were chosen from the 2002 TREC interactive track, and the remaining two from other work [5].

4. RESULTS

Ignoring the practice task, the 28 users submitted a total of 525 queries over 7 tasks. We did not see significant differences in the number of queries submitted to each engine.

During the experiment, users clicked on a total of 459 search results. Of the 525 queries, a total of 144 (33 %) queries did not have any clicks.

4.1 Overall success and satisfaction

Users appear to have been uniformly successful across all engines—users were successful around 80% of the time, and a χ^2 test indicated that the engine used caused no significant effect on reported user success. This continued success, even with degraded engines, is consistent with previous work.

A manual examination of the ratings revealed that users tended to rate queries highly after they were successful in a task. In order to prevent this affecting results, we examine

the lowest reported satisfaction per user per task or engine. This gives us a lower bound on the user's satisfaction during the completion use of a given engine.

Figure 4 shows the lower bound on user reported satisfaction on each engine. Engines that we expect to be of higher quality tend to be associated with queries leading to higher reported satisfaction. Using a χ^2 test to compare the degraded engine to an expected distribution of ratings (actually the observed distribution on the good engine) we find *(i:1,2,4,5,6,7,8,9,10)*, *(s:1,2,4,5,6,7,8,9,10)*, *(i:1,3,8,9)* and *(s:1,3,8,9)* to differ significantly from the good engine with $p < 0.05$, while the other two engines showed no significant differences.

4.2 Satisfaction on spam vs irrelevance

In total, there were 173 queries with irrelevant documents inserted, and 197 queries with spam documents inserted. We label queries with at least one spam document inserted *spam*; queries with at least one irrelevant document inserted *irrelevant*; and queries with no insertions—i.e. those directly from the search engine—are labelled *live*.

Figure 2(a) shows the distribution of ratings between the different type of queries. The ratings differ significantly across the three types of query (χ^2 tests, $p < 0.05$). As can be seen in the figure, live result sets are twice as likely to be marked highly satisfactory, whereas a higher proportion of irrelevant or spam result sets are given lower ratings.

To see how early in the result set this preference emerges, we examine satisfaction based on the type of the result at rank one. Interestingly, the trend is not the same as when the type of the whole set was considered (Figure 2(b)). There is no statistical difference between the distributions of ratings when the first document is a live search result and when the first document is an inserted spam document. However, there is a significant difference between having the first document as a spam or live result and having the first document as an irrelevant result. Having a spam document at rank one does not affect the rating of the query, but an irrelevant document at the same position is likely to be associated with a lower rating.

4.3 Clicks on inserted documents

During the course of the experiment, 459 search results were clicked. However, there were proportionally fewer clicks than impressions on spam and irrelevant results, while there were proportionally more clicks on the live results (χ^2 test, $p \ll 0.005$)—unsurprisingly, users were less interested in spam and irrelevant documents than in live search results.

There is a significant difference between the distributions of ratings of queries in which users clicked a spam document, and the queries in which users clicked an irrelevant document. Figure 3 shows that users rated the queries in which they had clicked on a spam document higher than those in which they had clicked on an irrelevant document.

Where irrelevant documents were inserted, there is no significant difference in ratings whether or not irrelevant documents were clicked, which indicates that the click on the irrelevant document does not affect user satisfaction.

However, where spam was inserted there is a significant difference between the distribution of ratings for queries where the spam was or was not clicked. Contrary to expectation, it appears that queries in which spam documents were clicked are actually rated higher than queries in which spam documents are shown but only live search results are clicked (t-test, $p = 0.05$), although the difference is small. This implies that reading a spam document is sometimes useful, although the presence of spam documents in the list is usually a bad thing.

4.4 Time cost of inserted documents

Low value documents may introduce other costs to the user, such as extra reading time. To model reading time, we calculated the duration of each query, defined as the time between subsequent query submissions (or between the final query in a session and an answer for the task). We found queries with spam or irrelevant documents inserted took significantly longer (mean 10 seconds) to complete than queries without insertions. However, there was no significant difference between the mean time taken on queries with spam and on queries with irrelevant documents.

We also calculated the mean time per click on each query, defined here as the duration of each query divided by (1 + the number of clicks). This mean time per click showed the same properties as the mean duration. The same is true again for the mean duration of queries with clicks on spam, irrelevant, or live results. We conclude that the insertion of spam or irrelevant documents causes the query duration to lengthen equally.

An alternative way to model reading time is to measure the time to first click on each query. Again, queries with no insertions require significantly less reading time (in this case around 2 seconds), and there was no significant difference between the spam and irrelevant queries.

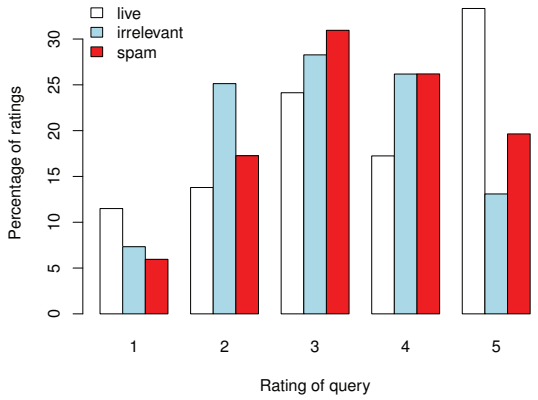
5. DISCUSSION AND CONCLUSIONS

Queries with no inserted documents achieve nearly twice the proportion of "extremely satisfactory" ratings as those with spam, and nearly three times that of those with irrelevant documents (Figure 2(a)). Interestingly, queries with spam documents inserted are generally rated as more satisfactory than queries with irrelevant documents inserted, and this is especially pronounced at rank one. This implies that live results are usually preferred to spam documents, and that both are always preferred to irrelevant documents. The result that both spam and live results at rank one receive the same overall satisfaction rating may imply that users have trouble telling the difference between spam and non-spam documents.

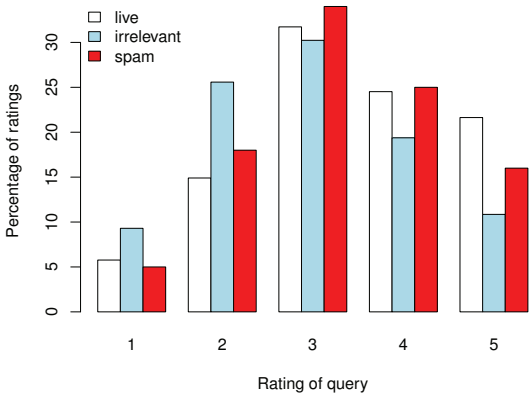
Both spam and irrelevant documents slow down users by around 10 seconds per query on average. Increased user effort—measured by time per query—has been shown to decrease user satisfaction scores [1], so this may explain some of the decrease in satisfaction caused by spam and irrelevant documents. However, because the time per query was increased identically between spam and irrelevant insertions, but satisfaction was lower when irrelevant documents were shown, it is unlikely that increased user effort is the only factor impacting on user satisfaction in this case.

It is especially interesting that while a click on an irrelevant document has no effect on user satisfaction, a click on a spam document does not make things worse and sometimes actually increases user satisfaction. This implies that many spam documents are actually useful, or at least attractive.

These observations suggest general principles. First, we should not rank simply by relevance, ignoring spam labels: given two documents of equal utility, the non-spam docu-



(a) Based on the type of document inserted



(b) Based on the type of document at rank 1

Figure 2: Ratings of queries based on the type of document inserted. 5 is high rating, 1 is low.

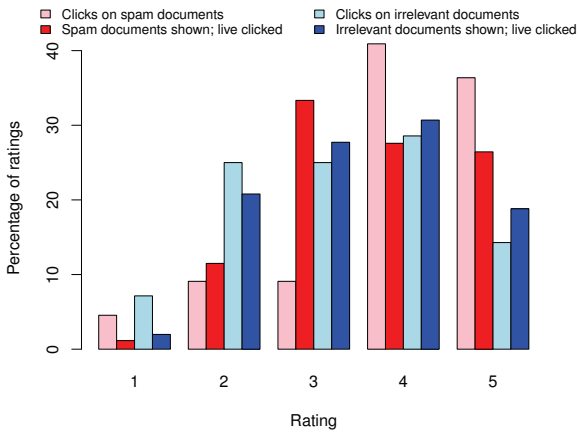


Figure 3: Ratings of queries in which spam or irrelevant documents received clicks.

ment should be preferred. On the other hand, nor can we discard all spam: spam documents might be relevant, and in many cases including spam doesn't hurt users' opinion overall. Ranking should consider both attributes.

One simple way to do this, especially if relevance and spam scores are quantised, might be to sort by relevance and break ties by spam scores. If estimated relevance and spam scores are used, this could provide a ranking for search engines; if documents are labelled for relevance and spam, this could provide a ranking against which to evaluate. Other more sophisticated ranking methods are of course possible. Regardless of the details, we argue that two scores, not just one, are needed for each document: both relevance and spamminess.

We summarise our observations of ranking and spam thus: The insertion of either irrelevant or spam documents into a result set increases the time a user takes to process the set, and by approximately the same amount; given two documents of equal utility, but differing spam scores, the document with the least spam should be preferred (live > spam); a result list without spam documents is more satisfactory than a result list with spam documents, assuming equal utility between the two lists; and an irrelevant document high in the result list is more damaging to user satisfaction than a spam document high in the result list (spam > irrelevant).

Relevant non-spam documents are better than spam; spam is better than irrelevance. Clearly, the design of a ranker or a metric to satisfy the above observations is a complex topic for future research, but at a minimum rankers should consider both relevance and spamminess as separate attributes.

6. REFERENCES

[1] A. Al-Maskari and M. Sanderson. A review of factors influencing user satisfaction in information retrieval. *JASIST*, 61(5):859–868, 2010.

[2] C. Castillo, C. Corsi, D. Donato, P. Ferragina, and A. Gionis. Query log mining for detecting polysemy and spam. In *Proc. WebKDD*. Springer, 2008.

[3] C. Castillo and B. D. Davison. Adversarial web search. *Foundations and Trends in Information Retrieval*, 4:377–486.

[4] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010.

[5] C. Liu, J. Gwizdka, and J. Liu. Helping identify when users find useful documents: examination of query reformulation intervals. In *Proc. IIiX*, 2010.

[6] Y. Liu, R. Cen, M. Zhang, S. Ma, and L. Ru. Identifying web spam with user behavior analysis. In *Proc. AIRWeb*, 2008.

[7] Y. Liu, M. Zhang, S. Ma, and L. Ru. User browsing graph: structure, evolution and application. In *WSDM (Late Breaking-Results)*, 2009.

[8] P. Thomas, T. Jones, and D. Hawking. What deliberately degrading search quality tells us about discount functions. In *Proc. SIGIR*, 2011.

[9] H. Yu, Y. Liu, M. Zhang, L. Ru, and S. Ma. Web spam identification with user browsing graph. In *Proc. AIRS*, 2009.