The usefulness of web spam

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

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

David Hawking
Funnelback Pty Ltd
Canberra, Australia
david.hawking@acm.org

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

Abstract *Spam comprises at least 60% of the public web, and search engine companies invest considerable effort in rejecting these apparently useless pages. But how bad are spam pages in search results? Can spam be dealt with as a side-effect of dealing with page utility, or is the relationship more complex?*

Thirty-four volunteer judges rated selected individual documents first on usefulness to a specified task and then on degree of "spamminess". Our results show that the relationship between spamminess and utility is far from clear cut; judges found that an important proportion of spam documents were useful. We conclude that evaluation should consider both utility and spamminess, as separate factors; and that search engines should not summarily discard spam pages but should take their utility into account as well.

Keywords User Studies Involving Documents; Web Documents

1 Introduction

Between 60 and 80% of pages on the web may be spam, as of 2009 [4]. Detecting spam documents on the web has therefore received much recent attention [2, 3]. However, relatively little attention has been paid to the problem of spam nullification—understanding how to correctly deal with spam documents once they have been detected [2, 6]. A common assumption is that spam pages should either be removed from the index or filtered out at query time, and never shown to users. However, this may be counter-productive if and when a spam page turns out to be useful, and both search engines and search evaluations should take this into account.

In order to understand how to correctly deal with spam documents, it is important to investigate the relationship between spam and utility. We conducted a labelling experiment to investigate whether there is a relationship between a page's usefulness and spam score. Judgements were made on a carefully selected pool of documents, which included known relevant documents, known irrelevant documents, and known spam documents.

Note that in this study we consider judge's ratings of individual documents in isolation. A companion study of the relative effect of spam and irrelevant documents on user satisfaction with search engine result pages is reported in [5].

Proceedings of the 16th Australasian Document Computing Symposium, Canberra, Australia, 2 December 2011.

2 Queries and documents

To build the pools, popular queries were extracted from the FAST search engine query log [8]. The list of queries was pruned to remove adult content and likely single answer navigational queries (such as "hotmail.com"), and then unique queries were ranked by frequency. Five queries that lent themselves to information gathering tasks were hand selected from the top 20 of the pruned list, and task statements were created for each of these queries. The queries were "free posters", "moon landing", "online tv", "recipes", and "dictionaries". Task statements are included in Appendix A.

For each query, we built a pool of nine documents.

- Two documents were chosen from the first page of results of each of two major search engines, for a total of four relevant documents.
- Three documents were chosen from later pages of search results. These were chosen to be irrelevant, but still to be plausible results—e.g., they still contained related terms.
- Two spam documents were retrieved from an index of pages from the UK-2006 collection which had been labelled as spam [3].

These ratios were chosen based on observations of commercial and our own search engines in previous experiment.

3 Judgements

Judgements were obtained in a two phase process. In the first phase, judges were presented with each query and the corresponding task description. Each document from the current pool was presented as if it were a search result, with title, query-biased summary, and a link to the document itself. Document ordering within each pool was randomised. Judges were then asked to rate each result on the scale {*very useful, useful, ok, not useful, totally useless*} according to how useful they believed it would be for completing the task. This five point scale is inspired by the scale used by Wu and Davidson [9], but we use the word *useful* instead of *relevant*. Judges were not given any instruction in what "useful" meant.

In the second phase of the experiment, we educated judges to assess pages for spam content, using the instructions presented to judges for the UK-2006 and UK-2007 web spam collections [3]. Judges were presented with each document from each result pool again, and asked to rate documents on the scale {*completely spam, borderline spam-leaning, borderline normal-leaning, normal*}. These labels were inspired by the labels from the UK collections, but we include two levels of borderline labels because of the anticipated disagreement on pages labeled *borderline* [3]. We did not include the *can-not-classify* label, because we ensured all results in the pools were available at the time of the experiment. Within-pool document ordering was again randomised.

34 volunteer judges, largely postgraduate students, were solicited from the university community. Judges were compensated with a movie ticket upon completion of their judgements. Some judges did not complete the judgement process, causing some of their judgements to only have a usefulness score. Removing these incomplete judgement pairs, the 34 judges created a total of 1123 ⟨*judge, document, usefulness score, spam score*⟩ judgements across the 45 documents.

This is a relatively high dropout rate: we have lost 27% of possible judgements. Judges worked remotely and unsupervised, so it is likely that some simply got bored. On the other hand, this suggests that those judgements we did collect are from motivated judges. Since documents were presented in random order, we do not expect any significant bias due to attrition.

Since our volunteers judged each page for utility before judging them for spamminess, there is a chance of some carry-over effect. Having decided a page is useful, and with a notion of spam as useless by definition, judges may be prone to labelling *very useful* or *useful* pages as non-spam where they would have labelled them spammy without the prior prompt. This may mean that there are even more useful spam pages than are counted here. In this respect, our conclusions are possibly conservative.

4 Results

Examining the spam scores for all the judgements at each level of utility produces Figure 1(a). The distribution of spam scores varies significantly across levels of utility (χ^2 test, $p \ll 0.0005$), indicating that spam scores vary somehow with utility; it is clear that in fact the more useful pages are less likely to be spam. A similar relationship is visible in Figure 1(b), where spam labels explain some of the difference in utility scores and spammier pages are less likely to be useful.

Note however an interesting minority in either case: when a document was judged *very useful*, in more than 20% of cases it was also judged *borderline-spam-leaning* or *completely spam*. In the case of judgements of *completely spam*, over 15% were also judged *useful* or *very useful*. Clearly, not all spam is useless and some is in fact useful.

Our judges did not tend to agree with each other in their labelling. We recorded Krippendorff's α of 0.249 for spam judgements, and 0.483 for relevance judgements.[1] This is low agreement, especially for "spaminess", which suggests that spam is hard to spot even when judges are given careful instruction. Castillo et al. note [3]

> "a common problem raised by the judges was that the evaluation of borderline cases is very subjective. Indeed, many Web sites that use spam techniques also provide some contents, so that it is very difficult to classify them as spammers."

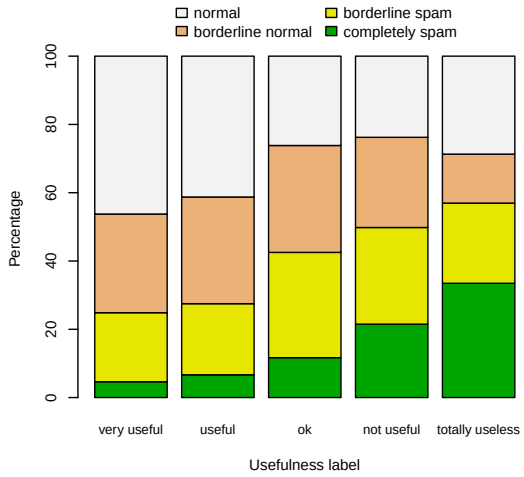
We agree. Inter-judge agreement is likely to be low for most experiments of this nature: Castillo et al. report Fliess's K of 0.56, "moderate" agreement, on a simpler problem with a three- (not four-) point scale and with only two judges per document (not up to 34).

5 Discussion

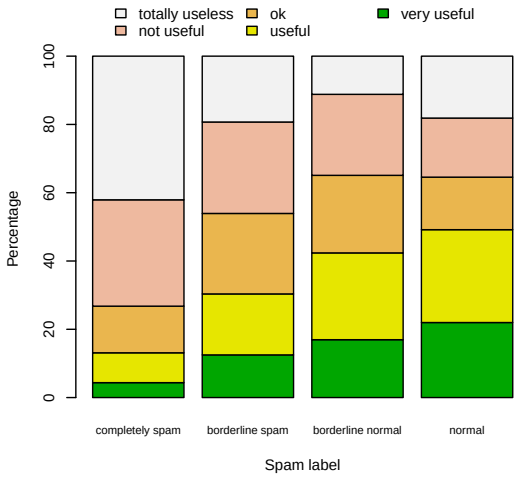
The percentage of *totally useless* documents that were labelled spam is nearly 60%, which is close to the background probability that a given page from 2009 was spam [4]. There appears to be no relationship between documents marked as *totally useless* and any spam label other than *completely spam*—as one would expect, just because a document is not spam does not mean it is useful.

Observing Figures 1(a) and 1(b), only a small percentage of *completely spam* documents are considered *useful* or better. Additionally, in Figure 1(b), we see considerably more documents marked as *totally useless* in the *completely spam* category than in any other spam category. These two observations imply it may be safe

[1] α, which measures inter-judge agreement, ranges from 0 (data are random) to 1 (judges are perfectly in agreement) and can be adapted to multiple judges, missing observations, and ordinal rather than categorical assessments [1, 7].



(a) Judgements of spamminess, conditioned on level of utility.



(b) Judgements of utility, conditioned on level of spamminess.

Figure 1: All 1123 ⟨ utility, spamminess ⟩ judgements.

to remove the spammiest documents from the index, as they are most likely not useful.

However, although there is a slight trend that useful documents have lower spam scores (Spearman's $\rho = -0.26$), it is difficult to make reliable assumptions about the potential spam score of a document which is a little bit useful; or about the usefulness of a document which is a little spammy. Importantly, more than 20% of *very useful* documents were labelled *borderline-spam-leaning* or *completely spam*.

Cormack et al. [4] found that their spam filter (used to produce the ClueWeb'09 spam labels) was also effective at detecting documents that had been labelled irrelevant by TREC assessors—although less effective than it was at detecting spam documents. This suggested that spam documents may be generally irrelevant, which is a stronger effect than we see here. Due to the larger sample used by Cormack et al., and our hand-crafted document selection method, their results may be biased towards including "spammier" spam examples. It is possible that a large fraction of real-world cases are both obviously spam and obviously useless (e.g. pages from a link farm, as seen in Figure 2(a)). Even if these pages do dominate, it is important not to ignore the potential usefulness of completely or borderline spam documents.

The low inter-judge agreement seen here and in the UK-2006 collection reinforces this: if we believe a page is spam, and remove it, not everyone will agree with the choices we make.

In related work [5], we have seen that users are sensitive to spam in result sets, so in general spam should be supressed. However, discarding all spam pages will mean that some useful pages are thrown away. Neither utility nor spamminess can be the only factor in a good ranking—search engines, and evaluation measures, must account for both.

What are the useful spam pages? A manual inspection of the pages that were labelled as *borderline-spam-leaning* or *completely spam* and with a usefulness score of *ok* or better indicated that these pages tended to contain excessive advertising but also provide some useful service such as a store or edited content. Many of these pages were much more sophisticated than a simple copy of Wikipedia content. Figure 2(b) shows one such page, where users located in the US can obtain streaming television (although our judges were located in Australia, they still scored the page "useful" on average. Use of a proxy located in the US confirms that streaming television is available).

6 Conclusions

Our judges found that an important proportion of spam documents (around 13% to 40% depending upon the spam rating) were either *useful* or *very useful* to the specified task. This suggests that pages classified as spam should not be summarily excluded from search engine indexes. Since there is no clear-cut relationship between spamminess and utility, ranking algorithms and quality evaluation campaigns for Web search should take into account both utility and spamminess; neither alone will suffice.

(a) Example *completely spam* and *totally useless* page (averaging ratings across all users). This page offers no content itself, not even copied from elsewhere, so it is not useful; it serves only for advertising.

(b) Example *borderline-spam-leaning* but *useful* page (averaging ratings across all users). Although the page is heavily laden with advertising, and appears to be spam content, streaming television is available in one click.

Figure 2: Sample spam pages.

References

[1] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, Volume 34, Number 4, pages 555–596, December 2008.

[2] Carlos Castillo and Brian D. Davison. Adversarial web search. *Foundations and Trends in Information Retrieval*, Volume 4, pages 377–486.

[3] Carlos Castillo, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi, Massimo Santini and Sebastiano Vigna. A reference collection for web spam. *SIGIR Forum*, Volume 40, Number 2, pages 11–24, December 2006.

[4] Gordon V. Cormack, Mark D. Smucker and Charles L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. `http://arxiv.org/abs/1004.5168`, 2010.

[5] Tim Jones, David Hawking, Paul Thomas and Ramesh Sankaranarayana. Relative effect of spam and irrelevant documents on user interaction with search engines. In *Proc. CIKM*, 2011.

[6] Timothy Jones, Ramesh Sankaranarayana, David Hawking and Nick Craswell. Nullification test collections for web spam and SEO. In *Proc. 5th Int'l Workshop on Adversarial Information Retrieval on the Web*, 2009.

[7] Klaus Krippendorff. *Content Analysis, An Introduction to Its Methodology*. Sage Publications, California, USA, second edition, 2004.

[8] Amanda Spink, Seda Ozmutlu, Huseyin C. Ozmutlu and Bernard J Jansen. US versus European web searching trends. *SIGIR Forum*, Volume 36, pages 32–38, September 2002.

[9] Baoning Wu and Brian D. Davison. Undue influence: eliminating the impact of link plagiarism on web search rankings. In *Proc. SAC*, 2006.

A Task statements

The five tasks allocated were:

1. ("Free posters") You'd like to download and print some posters to decorate the walls of a teenager's bedroom. Please rate each result considering how useful it is for finding **free posters**.

2. ("Moon landing") July the 21st was the 40th Anniversary of the Apollo 11 Moon landings. With the recent coverage, you're interested in a summary of the events of the Moon landing. Please rate each result considering how useful it is as a summary of the Apollo 11 **moon landing**.

3. ("Online TV") You've heard that you can watch TV online, instead of using a television. Please rate each result considering how useful it is for watching **TV online**.

4. ("Recipes") You are looking for a site with a large collection of recipes to add to your home cookbook. Please rate each result considering how useful it is for finding **recipes**.

5. ("Dictionaries") You are looking for English to non-English dictionaries, so that you can file them away for later use. The more non-English languages, the better. Please rate each result considering how useful it is for finding bilingual **dictionaries**.