

A Framework for Measuring the Impact of Web Spam

Many web site operators strive to improve the layout and content of their sites using Search Engine Optimisation (SEO). However, some operators attempt to fool the ranking algorithms used by web search engines, using techniques commonly referred to as black hat SEO or web spam.

“[Web spam is] any deliberate action that is meant to trigger an unjustifiably favourable relevance or importance for some web page, considering the page’s true value”

Z. Gyöngyi and H. Garcia-Molina (2005)

Literature on the web spam problem has primarily focused on the classification of spam web pages. By combining classification techniques, spam pages can be detected with a high degree of precision, usually around 80%. However, it is unclear how much effect web spam has on the quality of results.

We use the UK2006 web spam collection, which is a web snapshot crawled without any spam rejection. The collection contains around 80 million pages, with roughly 4.16 billion links to and from 11,000 hosts. The collection includes human provided host-level labels from one of {“normal”, “borderline”, “spam”, “can not classify”} for 2,725 of these hosts.

Indexing and query processing

We indexed the 2 terabytes of the UK2006 collection on a single low cost machine. with 2 Intel P4 3.0GHz CPUs, 3GB of RAM, and just over 1 terabyte of disk space. The total system cost was approximately \$2,000 AUD. We used PADRE (a search engine built by CSIRO) for indexing and query processing. The total time to decompress and index the entire collection was 166.3 hours, resulting in a 129.5 GB index.

Does web spam affect quality?

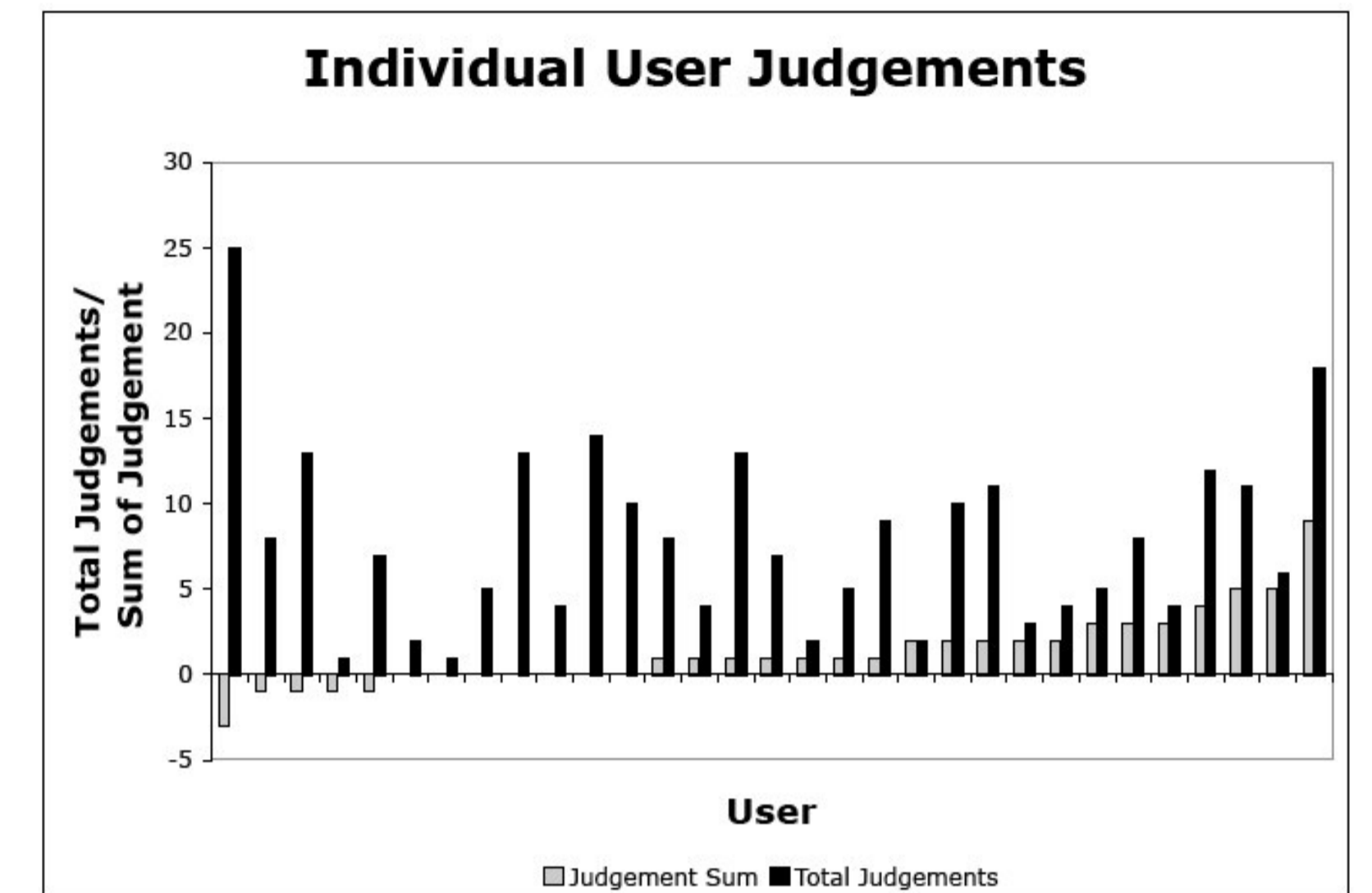
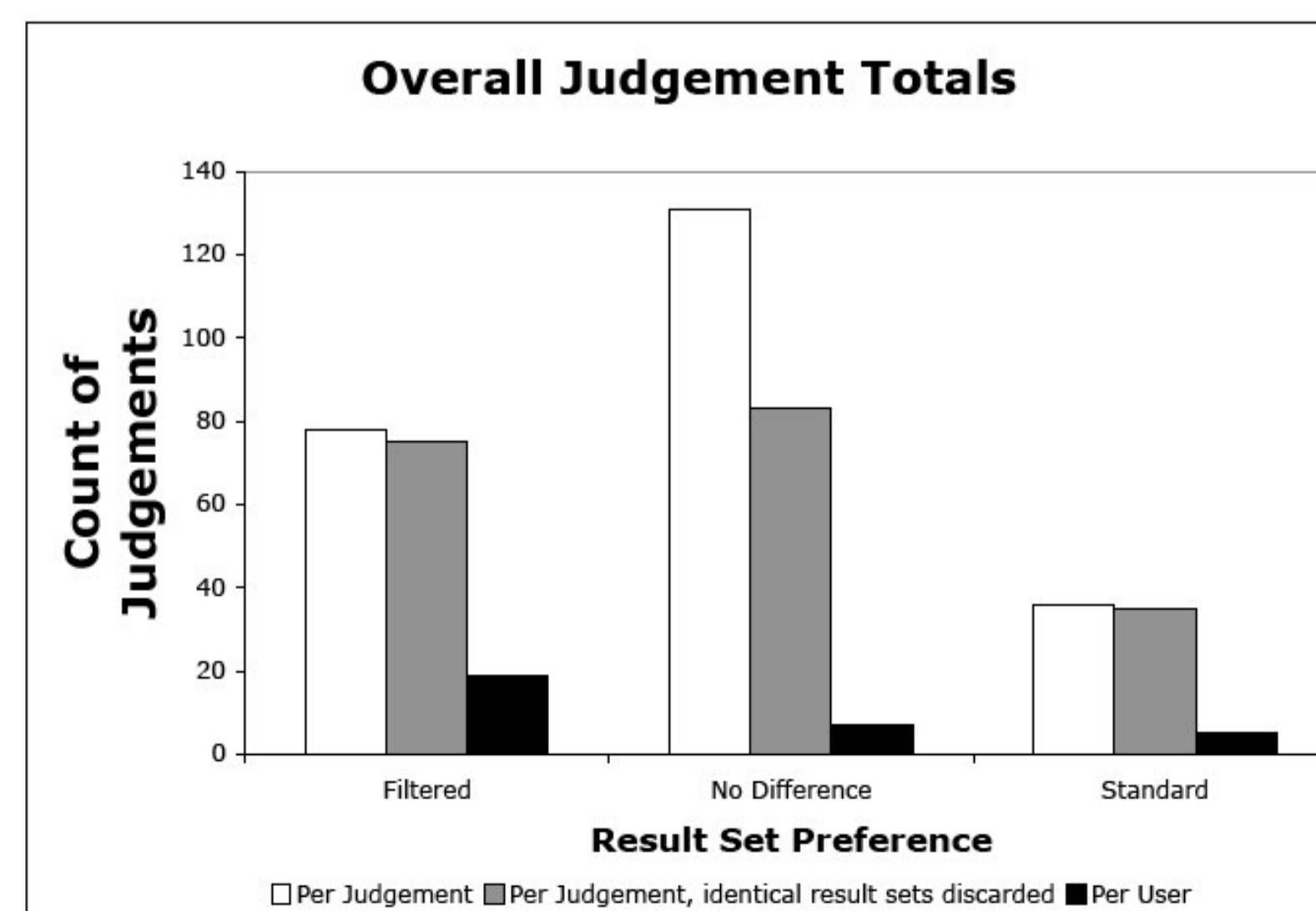
An easy treatment of web spam is to simply remove it from the result list. This is attractive because it enables the easy combination of spam detection techniques and because indexes do not have to be rebuilt.

Volunteer subjects submitted queries of their own using our two-panel evaluation interface, provided by Paul Thomas. We presented two sets of results for each query in random left-right order: *standard* comprises the first 10 results from our search engine; and *filtered* comprises the first ten after pages labelled as spam sites were removed. Users were invited to judge one list as being better than the other, or “no difference” between the lists.

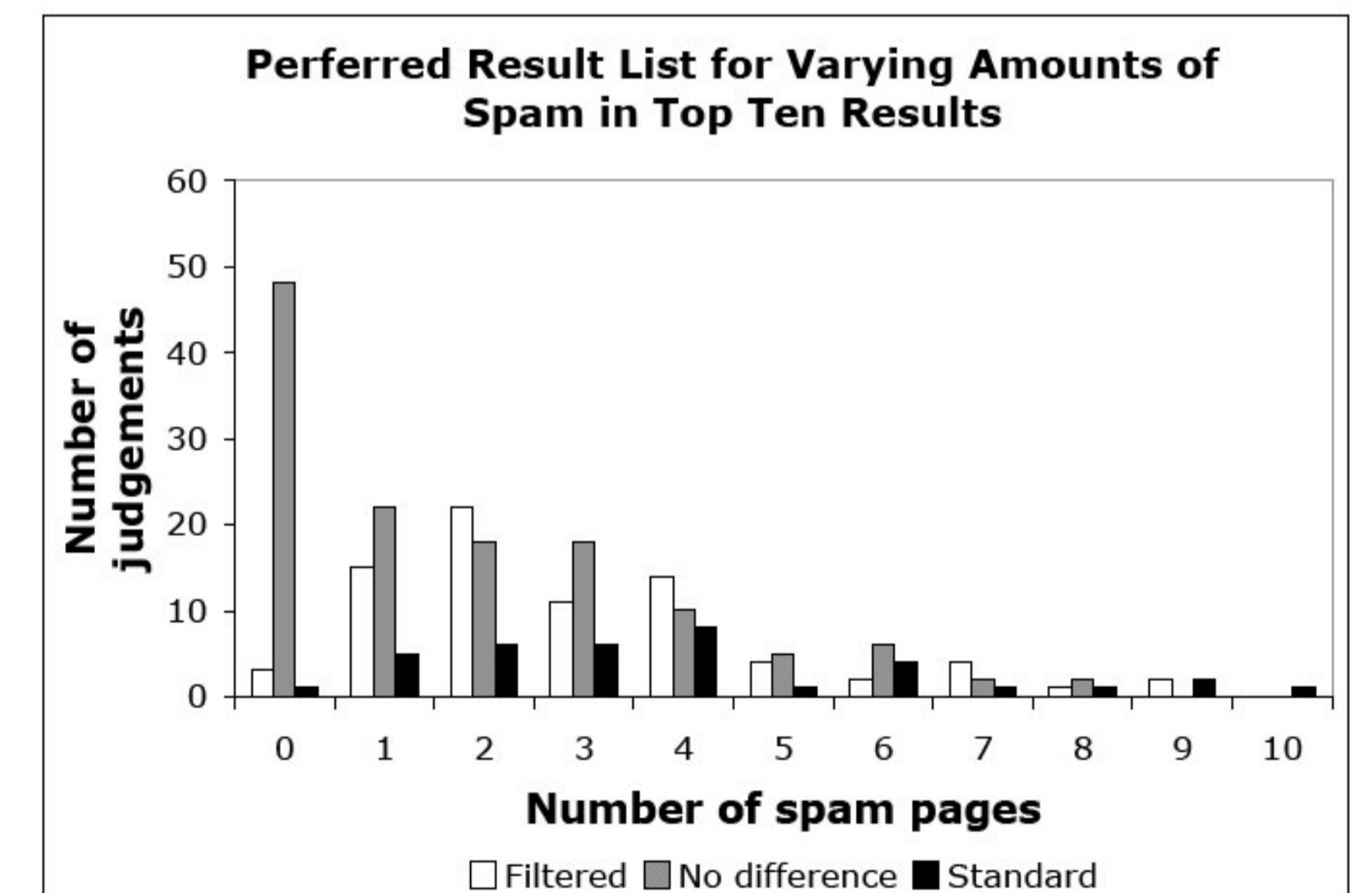


Results

There is a strongly significant difference between the total votes for filtered and standard. (Pearson’s chi-square test, $p < 0.0001$).



The black lines above show the total number of judgements, while the grey lines show the sum of that users judgements (plus one for each filtered vote, minus one for each standard vote). There appears to be no correlation between number of judgements made by a user, and judgement preference.



It is interesting that a stronger preference for the filtered set does not develop as more spam appears in the results.