Does brandname influence perceived search result quality? Yahoo!, Google, and WebKumara

Peter Bailey
CSIRO ICT Centre
Canberra, Australia
peter.bailey@csiro.au

Paul Thomas
Australian National University
Canberra, Australia
paul.thomas@anu.edu.au

David Hawking
CSIRO ICT Centre
Canberra, Australia
david.hawking@csiro.au

Abstract *Improving the quality of search engine results is the goal of costly efforts by major Web search engine companies. Using* in situ *side-by-side result set comparisons and random assignment of brandnames to result sets, we investigated whether perceptions of quality were influenced by brand association. In the first experiment (15 searchers) we found no significant preference for or against results labelled "Google" relative to those labelled "Yahoo!". In the second experiment (20 searchers) result sets were again generated by Google and Yahoo! but were randomly labelled "Yahoo!" or "WebKumara" (a fictitious name). Again, we found no significant preference for one brandname label over the other. Contrary to previous findings, we found a statistically significant preference for Google-generated results over those of Yahoo! when data from three separate experiments (total 70 subjects) was combined.*

Keywords Information retrieval

1 Introduction

The quality of search results is vitally important to the success of major Web search engines. It is understood that the operators of these search engines continually monitor their result quality and that of their competitors. However, over recent years, the market share of searches carried out by users on different search engines has changed. Recent reports credit Google with an increasing market share [4] at the expense of Yahoo! and MSN in particular,[1] although there does not appear to be a strong actual difference in result quality [6]. Other factors, such as a difference in *perceived* quality, may be contributing to this difference in market share.

2 Related work

Objective Web search engine evaluation is known to be a difficult task, due to the inability to compare individual search engines against a common corpus. Hawking et al. compared twenty public search engines using TREC Web track-based methods using a set of 54 queries from real Web search engine logs [1]. They compared 20 different Web search engines across seven different standard IR measures including P@n ($n < 20$), MRR1, relative recall, and average precision.

Tang and Sun rejected precision/recall metrics as being inappropriate, and instead investigated a small number of new user effort sensitive evaluation measures, from a set of 8 topics from 4 PhD students in [5]. This study, while interesting, may have insufficient topics to be reliable. (It is widely accepted in the IR community that typically 50 or more topics are required for such investigations [3], though the exact number depends on the statistical power required, the size of the effect to be observed and the number of variables present in the experimental design.)

Jansen et al. [2] have investigated branding, to understand why some search engines have large market share despite result sets being of similar quality. To control result quality, one set of results was generated for each of four queries; these result sets were branded with logos and other design elements from Google, Yahoo!, MSN, and AI^{2}RS to give a total of sixteen "result sets". Participants were asked to judge the relevance of each result. Sets branded with Yahoo! had highest judged "average precision", despite being identical to the others, and Google, MSR, and AI^{2}RS followed. Jansen et al. used only four queries however and did not report on the statistical significance of their results.

Thomas and Hawking [6] present a new method for allowing in-context judgements by real users (and their real information needs and search queries) with side-by-side comparisons between different systems. They used this method to report on user perceptions of

[1] Note that in markets such as China and Korea local engines are reported to dominate.

Proceedings of the 12th Australasian Document Computing Symposium. Melbourne, Australia, December 10, 2007.

result quality on two anonymised whole-of-Web search engines. Users were asked to use the search interface for their normal day-to-day search needs, and to judge which of the two result lists was preferred or if no difference could be detected. Judgements, including of "no difference" were not compulsory. Unlike the method of Jansen et al., Thomas and Hawking were able to use naturally-arising queries and to compare complete result sets rather than judging individual documents; however, they did not consider the effect of branding.

3 First branding experiment

We replicated the Thomas and Hawking experiment and method: for each user-generated query, two result sets were presented side by side in a random order and users were given the opportunity to indicate which (if either) was "better". "Better" was not further defined. Instead of anonymising the Web search engines as in the Thomas and Hawking work, we identified the individual result lists with the brand names of two major search engine companies. One list was labelled "results from Google" and the other "results from Yahoo!"; however, the labels were assigned randomly so that they only reflected the true source of the results 50% of the time. The experimental interface is illustrated in Figure 1.

Note that results are branded only with a name, rather than for example colours and logos. Result colour layout among all the major search engines has consolidated around a white background, blue title hyperlinks, black snippet text, and green url and meta information.

19 users participated, of whom 15 submitted at least one usable query. (A usable query is one where a preference is expressed, even if it was "no difference".) In total the users submitted 370 queries and expressed 123 definite preferences. A definite preference is where a user judged one list to be better than another; no preference is where the "no difference" judgement is made, or no selection is made at all. Since the experiment aimed to find out whether a difference existed, only definite preferences were considered as positive evidence. Participants' demographics are summarised in Table 1. Participants had a varied range of educational disciplines and employment positions and organisations, but there is a bias towards postgraduate education levels.

Results are shown in Table 2. Note that the numbers in the results table are the number of people, not the number of judgements, since we wish to establish the number of individual users who had a discernible preference one way or the other over the totality of their own searches (not on any individual search carried out).

Sex	Male: 10, female: 9
Education	Postgraduate degree: 13, first degree: 5, other: 1
Age	26–68 (mean 34.9, std. dev 8.4 years)

Table 1: User demographics for the first experiment.

Our analysis considers both the real and the labelled provider of the preferred result sets. These may be different: for example, a participant who expressed preferences for 5 result sets from Google and 2 from Yahoo! will be recorded as preferring results from Google overall. Since labels are applied at random, these seven preferred sets may have been labelled with "results from Google" in only 1 case and "Yahoo!" in the other 6; this will be recorded as a preference for results labelled "Yahoo!".

If a participant had an equal number of result sets marked from each provider or label, no overall preference was recorded. These were treated conservatively, as evidence against both alternatives.

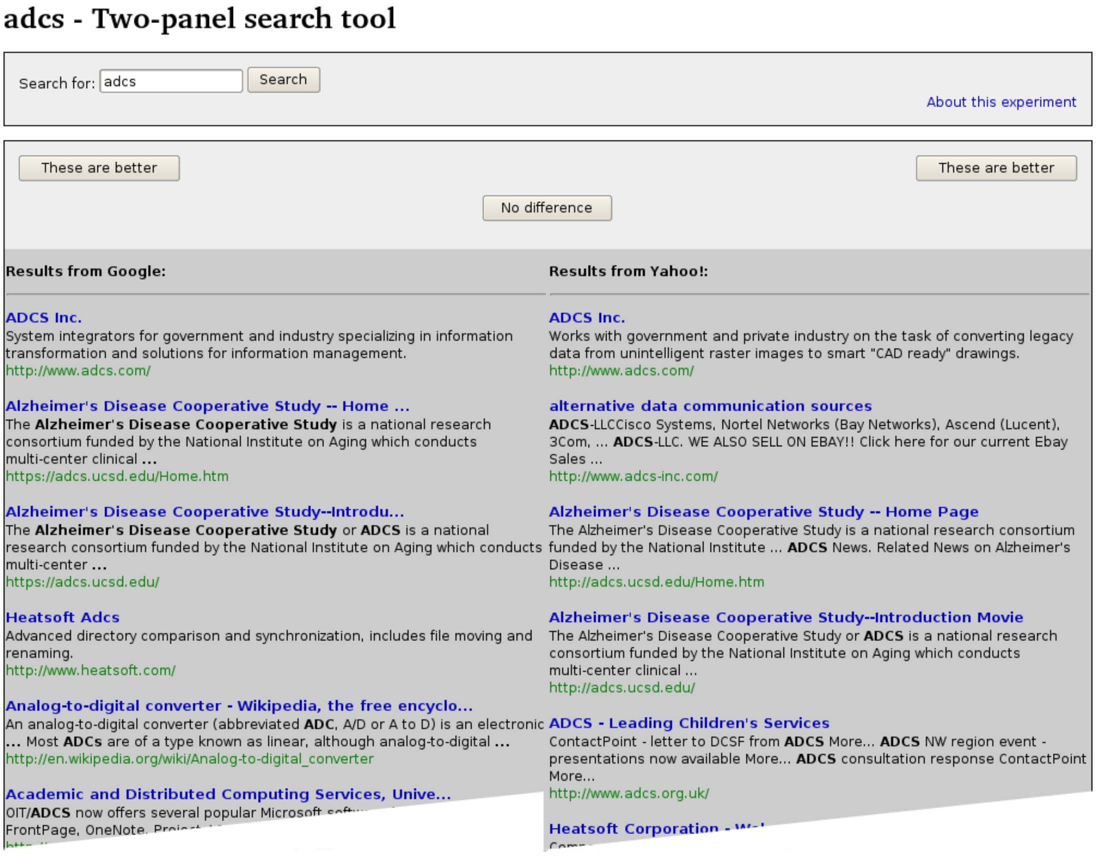
We used both binomial sign tests (less powerful) and Wilcoxon signed-rank tests (more powerful) to assess the branding results. No significant difference was found with either test. We conclude that our participants were not influenced by the brand reputation of either search engine.

4 Second branding experiment

Given this result, we decided to conduct a second experiment to see whether users would prefer the results branded as coming from a well-known search engine (Yahoo!) to the results branded as coming from an unknown search engine. We invented the name of a search engine — "WebKumara" — for this purpose. However, in all other respects the experimental method remained unchanged. Thus results were provided by both Google and Yahoo! and randomly branded as coming from either WebKumara or Yahoo!, and the quality of search results was unchanged.

20 users participated, submitted 284 queries, and expressed 115 definite preferences. Results are shown in Table 2. (User demographics were similar to those in experiment 1.)

Binomial sign tests and Wilcoxon signed-rank tests were performed on these results. Neither test revealed a significant difference between results labelled "WebKumara" and those labelled "Yahoo!". Searchers do not seem to prefer result rankings associated with a well-known search engine over those from an unknown one.



adcs - Two-panel search tool

Results from Google:

ADCS Inc.
System integrators for government and industry specializing in information transformation and solutions for information management.
http://www.adcs.com/

Alzheimer's Disease Cooperative Study -- Home ...
The **Alzheimer's Disease Cooperative Study** is a national research consortium funded by the National Institute on Aging which conducts multi-center clinical **...**
https://adcs.ucsd.edu/Home.htm

Alzheimer's Disease Cooperative Study--Introdu...
The **Alzheimer's Disease Cooperative Study** or **ADCS** is a national research consortium funded by the National Institute on Aging which conducts multi-center **...**
https://adcs.ucsd.edu/

Heatsoft Adcs
Advanced directory comparison and synchronization, includes file moving and renaming.
http://www.heatsoft.com/

Analog-to-digital converter - Wikipedia, the free encyclo...
An analog-to-digital converter (abbreviated **ADC**, A/D or A to D) is an electronic **...** Most **ADCs** are of a type known as linear, although analog-to-digital **...**
http://en.wikipedia.org/wiki/Analog-to-digital_converter

Academic and Distributed Computing Services, Unive...
OIT/**ADCS** now offers several popular Microsoft soft
FrontPage, OneNote, Proj

Results from Yahoo!:

ADCS Inc.
Works with government and private industry on the task of converting legacy data from unintelligent raster images to smart "CAD ready" drawings.
http://www.adcs.com/

alternative data communication sources
ADCS-LLCCisco Systems, Nortel Networks (Bay Networks), Ascend (Lucent), 3Com, ... **ADCS**-LLC. WE ALSO SELL ON EBAY!! Click here for our current Ebay Sales ...
http://www.adcs-inc.com/

Alzheimer's Disease Cooperative Study -- Home Page
The Alzheimer's Disease Cooperative Study is a national research consortium funded by the National Institute ... **ADCS** News. Related News on Alzheimer's Disease ...
http://adcs.ucsd.edu/Home.htm

Alzheimer's Disease Cooperative Study--Introduction Movie
The Alzheimer's Disease Cooperative Study or **ADCS** is a national research consortium funded by the National Institute on Aging which conducts multi-center clinical ...
http://adcs.ucsd.edu/

ADCS - Leading Children's Services
ContactPoint - letter to DCSF from **ADCS** More... **ADCS** NW region event - presentations now available More... **ADCS** consultation response ContactPoint More...
http://www.adcs.org.uk/

Heatsoft Corporation - W

Figure 1: User interface for the first experiment.

First experiment (19 participants; 15 users with preferences)			
Preferred results from Google:	9 users	Preferred results labelled “Google”:	7 users
Preferred results from Yahoo!:	3 users	Preferred results labelled “Yahoo!”:	4 users
No overall preference:	3 users	No overall preference:	4 users
Second experiment (20 participants; 20 users with preferences)			
Preferred results from Google:	13 users†	Preferred results labelled “WebKumara”:	12 users
Preferred results from Yahoo!:	2 users	Preferred results labelled “Yahoo!”:	6 users
No overall preference:	5 users	No overall preference:	2 users
Aggregated Google v. Yahoo! results (70 users)			
Preferred results from Google:	45 users‡		
Preferred results from Yahoo!:	17 users		
No overall preference:	8 users		

Table 2: Overall user preferences. † significant at $\alpha = 0.05$, Wilcoxon signed-rank test. ‡ significant at $\alpha = 0.01$, same test.

These results contrast with those of Jansen et al. [2], who saw a preference for major search engine brands. We note that there are differences in methodology which may explain the difference: while Jansen et al. asked for judgements on each individual document, we asked for judgements on entire document sets. Users may consider one search engine to be better at returning relevant, but essentially duplicate, documents; this will favour the search engine on their measure but not on ours.

5 Preference between Yahoo! and Google

We aggregated data for preference between actual engines from the first and second branding experiment with those reported in the Thomas and Hawking experiment [6]. Note that if the same user participated in more than one experiment their results were combined unless they used different user-ids (which to preserve privacy would be undetectable by the experimenters).

The bottom panel of Table 2 shows that the consistent small advantage to Google across the three experiments translates to a highly significant ($p < 0.01$) preference across the aggregated users. Caution must be exercised in interpreting this result:

- 25 out of 70 users did not find Google results to be better.
- Data was collected over a period of many months during which time ranking functions and index contents may have varied considerably.
- There are significant biases in the demographics of our users — they tended to be well-educated Australian and British people.
- Although the result is highly significant, the size of the effect (of a preference for Google) may be small.

6 Conclusions

In judging result quality, users do not appear to be strongly influenced by the brandname of the search engine alleged to have generated the results, even if that brandname is totally unknown (in fact, fictitious). However, in everyday use their choice of search engine may be influenced by yet other factors besides result quality, such as other aspects of branding (e.g. colours and logos), speed of delivery of a page of results, effectiveness of advertising, or pre-loading as the default search engine in computer or browser setup.

In each of our individual experiments we found a small preference for results generated by Google over those generated by Yahoo! at the times when the data was collected. Combining all the data we found a highly significant difference in favour of those from Google, with 64% of participants exhibiting a (blind) preference for the engine with the largest market share. This result should be interpreted with caution, taking into account the caveats expressed in the previous section.

Acknowledgements We thank the many participants who carried out judgements over their queries.

References

[1] David Hawking, Nick Craswell, Peter Bailey and Kathleen Griffiths. Measuring search engine quality. *Information Retrieval*, Volume 41, Number 1, 2001.

[2] Bernard J Jansen, Mimi Zhang and Ying Zhang. Brand awareness and the evaluation of search results. In *Proc. WWW*, 2007. Poster.

[3] Mark Sanderson and Justin Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proc. ACM SIGIR*, 2005.

[4] Danny Sullivan. comScore media metrix search engine ratings. `http://searchenginewatch.com/showPage.html?page=2156431`, August 2006.

[5] Muh-Chyun Tang and Ying Sun. Evaluation of web-based search engines using user-effort measures. *Library and Information Science Research Electronic Journal*, Volume 13, Number 2, 2003.

[6] Paul Thomas and David Hawking. Evaluation by comparing result sets in context. In *Proc. CIKM*, 2006.