Are People Biased in Their Use of Search-Engines ?

Keane, M.T., O'Brien, M. & Smyth, B.

Adaptive Information Cluster, School of Computer Science & Informatics, University College Dublin, Ireland

Search-engines are among the most used resources on the Internet. Google [3], for example, now hosts over eight billion items and returns answers to queries in a fraction of a second, thus realising some of the more far-fetched predictions envisioned by the pioneers of the World Web Web [2]. In the present study, we assess whether people are biased in their use of a search-engine; specifically, whether they are biased in clicking on those items that are presented as being the most relevant in the search engine's result list (i.e., those items listed at the top of the result list). To test this bias hypothesis, we simulated the Google environment systematically reversing Google's normal relevance-ordering of the items presented to users. Our results show that people do manifest some bias, favoring items at the top of result lists, though they also sometimes seek out high-relevance items listed further down a list. Later, we discuss whether this bias arises from people's implicit trust in a search engine, like Google, or some other effect.

Introduction

The World Wide Web provides access to an unparalleled volume of information at time costs that are orders of magnitude lower than those required for traditional media. The critical jump off point for this vast repository is typically provided by the results returned by search engines to short, user queries: Google [3], for example, returns results to an average of 200 million queries everyday, queries that are typically about two words long. Like many search engines, Google uses the collective intelligence of the web to rank-order pages of relevance to a particular query. Each page in this ordered list is typically summarised by a clickable title, some snippets of the page's content (with highlighted matching content words) and a web-address link.

Rational searchers should assess each of these page summaries against their information need and click on the one that seems to be the most relevant. However, people may not search in such a rational way. They may manifest biases; for example, they might click one of the top-listed results without much checking against their information need. Such biases, if they exist, could be due to users coming to implicitly trust a search engine. That is, over time, as a search engine consistently delivers relevant pages towards the top of its result lists, users might come to assume that the top results are indeed the best (see [4], [5], [9] and General Discussion).

In this paper, we present a study that assesses whether people manifest a search bias in their use of one search engine, Google. We simulated the Google environment controlling the results given to user queries. The key manipulation was to compare users' responses when they received result-lists in their normal ordering versus a systematically-reversed order. If people are biased in their search then they will not notice that the relevance rankings have been reversed. That is, they should respond identically to the normal and reversed lists clicking on results placed first at the top of the lists. If they are not biased then they should respond differently to the normal and reversed lists; specifically, they should hunt down the reversed list to find highly-relevant items listed last. To presage our results, the truth seems to lie somewhere between these two extremes. There is definite evidence of bias in people's Google searches; they tend to click on first-listed items, though they sometimes will seek out highly-relevant, bottom-listed ones too.

Method

Thirty Science undergraduates at University College Dublin's were paid to participate in the study. Participants were asked to answer 16 questions on Computer Science (e.g., "Who invented Java?") by running as many queries as they liked on the simulated Google environment. The interface was designed to have the look and feel of Google. Indeed, all reported that they really thought it *was* Google. The simulated system was built using comprehensive search logs from a live user trial in which a separate group were asked to answer the same 16 questions [7]. All of the queries used by this group were stored, as were all of the result-lists returned to these queries. We then created a database linking specific queries to their appropriate result-list.

This database sat behind our simulated Google interface and was used to systematically counterbalance the presentation of result-lists in either a normal or reversed ordering when a user entered a query.

The results-lists returned to a given query were presented in either their original relevance ordering or a reversed ordering in a counterbalanced way across trials of the experiment. The order in which questions were presented was randomised for each participant to counteract any learning effects that might occur in the course of a trial. Participants were instructed to limit the number of query terms to two or less, to parallel typical user behaviour (i.e., the average web user uses enters between 1 and 3 query terms [8]). For each question we recorded the number and names of the queries entered and the search results clicked on by users. The timing of each transaction was also recorded. Participants were asked to complete a form detailing their answers to the questions and sessions averaged 1.5 hours.

We also carried out a ranking post-test to see whether people agree with Google's relevance ordering of results. This post-test was carried out on a sample of the result-lists using a new group of 14 students. These participants were asked to manually rank the presented result-lists from the search experiment (on a 1-10 scale from "most likely to select" to "least likely to select"). A sample of 16 result-lists from experiment was used, based on those result-lists returned to the most frequently-used query for each of the 16 questions. So, this sample should cover those result-lists that contribute most to any effects found in the experiment. Each participant received the result lists in a randomised order and the results in each list were also randomised for every participant. This procedure was adopted to ensure an accurate assessment of people's relevance ranks, independent of any possible bias effect. People took an hour to complete this ranking task during which participants only completely ranked a subset of the presented result-sets.

Results & Discussion

The dependent measure was "first clicks"[6], the first chosen link by a given user in a returned result list to a given query. The data were analysed in a 2 (condition; normal versus reversed) x 10 (relevance-rank; 1-10) design treating

queries as the random factor. That is, for each query we recorded the proportion of people that chose a particular ranked-result, noting whether this occurred in a list that was normal or reversed. The two-way analysis of variance (ANOVA) with condition and relevance-rank revealed a main effect of relevance-rank [$F(9,319) = 102.14$, $p<0.01$, $MSe= 0.89$], and a reliable interaction between the condition and relevance-rank [$F(9,319)=11.31$, $p<0.01$, $MSe=0.10$]. Tukey's post-hoc comparisons of the interaction showed that there were reliable differences between the first-click frequencies for the 1st, 9th and 10th relevance-ranks (see Figure 1).

These results clearly indicate that people's first-clicks in the normal and reversed conditions is not identical, providing evidence that people are "partially biased" in their search. Items with the highest-relevance ranks (i.e., items ordered first by Google) are chosen 70% of the time in the normal condition, but this rate drops to 10% in the reversed condition. In contrast, the 9th and 10th relevance-ranked items are chosen more often (13% and 41%, respectively) in the reversed condition than in the normal one (2% and 2%, respectively). Intermediately ranked items are much the same across both conditions.

The significance of what is happening here is readily apparent if one pictures the data by-position in the results lists (see Figure 1). This figure shows us that when lower relevance-ranked items are positioned first and second in the result list (as they are in the reversed condition) then they are being chosen more often by users, despite their limited relevance. In contrast, when the highest-relevance items are positioned last in the result list (in the reversed condition) they are being chosen considerably less often. In short, users are, in part, misled by the presented order of the items. However, sometimes people still hunt out the highest relevance item even when it is at the very bottom of the returned list.

Finally, the post-test showed that there is close agreement between people's rank of returned results and those posed by Google. People's mean rankings of the sampled result-lists correlate highly with the search engine rankings ($r^2 = 0.9124$ $t=9.13$; $df=8$; $p<.0001$). This result shows us that the items Google presents as the best are considered by people to be the best too. It is interesting that this finding occurs even when people have been given the result-lists in a randomly re-ordered form, as it

suggests that highly-relevant items in each result list were easily identifiable. This post-test also sheds some light on another issue to do with the relevance topology of the result-lists. One worry about the evidence is that the first 10 results in each list are roughly equal in relevance and that you only really start to get real relevance differences when you get to the 100^{th} or 200^{th} ranked items. If this were the case then the search behaviour observed would really only apply to result-lists with flat, relevance topologies. This concern is partly answered by the correlation reported above, but not fully. To get a better idea of the actual relevance topology we analysed the rankings produced by people in the post-test in a different way. For each of the 16 result lists sampled, we noted the mean rating given by people to each result in the list. If the relevance topology is flat for these lists then these mean ratings should all be roughly equal (recall, order effects are controlled for this data by randomisation). However, this is not what we found. There are a huge variety of different topologies for the results in each list; a few have a single highly-relevant item (with a mean rank of 1 or 2), others have several results given high mean ranks, while others have a linearly increasing relevance topology. This finding suggests that our random selection of questions for the experiment have generated a random selection of different relevance topologies, that are presumably representative of the topologies generated by Google. Furthermore, they are not all flat but hugely varied.

General Discussion

The present study clearly shows that people are "partially biased" in their search behaviour using Google. While it is known that people have a fondness for items at the beginning of written lists, the novelty of the present study is that it demonstrates such effects in the search-engine context in a systematically controlled way (i.e., through our forward-reversed paradigm). So, given that we have evidence of such a partial bias in search engines, the really hard question to answer is why ?

Recently, Joachims et al. [9] have using an eye-tracking paradigm in a similar study finding parallel effects to ours, that they interpreted as being due to people's development an implicit trust in search engines. That is, search engines could misleadingly over-promote an initially popular page because, having placed it at the top of the result list, it is clicked on unthinkingly by users, in turn increasingly the

likelihood of it being placed first, being clicked on unthinkingly by users and so on (see also [1],[4],[5]). This problem obviously applies to search engines that rely on histories of previous user choices (e.g., Smyth and I-SPY), but it could also apply to those using link-analysis schemes because the top-of-the-list pages are more likely to end up as the chosen link on people's web pages. If the bias found here is due to trust then search-engine designers may need to design systems to overcome such effects (for some solutions see Cho et al. [4]), Joachims et al [9]). However, we believe that this evidence and ours do not conclusively demonstrate that these biases are due to trust. The parsimonious interpretation is that these findings show a preference for first-placed items. To conclusively demonstrate trust, one would have to show an increase in the bias over time when some blank-slate user first comes to use a search engine. Given the difficulties of finding some novice users in the adult population it is clear that this is a challenging test to carry out.

Another possibility is that the bias largely a function of interactions between the perceived relevance of item-results to one's information need and some trade-off in terms of effort expended searching down the list. Some work has been done showing that such interactions occur, in cases where a list of items has a "peaky" relevance topology; that is, there is one highly-relevant item surrounded by obviously less-relevant items (see Howes [x]). Our findings seem closer to this type of search behaviour in that we only find a partial bias; people do sometimes search to the bottom of the list to find the highly-relevant items. An intriguing puzzle for future research is to determine the exact conditions under which people abandon biased behaviour. The work by Howes and his colleagues suggest one set of conditions, namely when it is clear from the relevance topology. However, our post-test shows that these exact conditions do not hold in the cases where people avoid the search bias. So, other factors must come into play as well.

Whatever the truth, it is clear that future information delivery systems have much to learn from such detailed analyses of user search behaviour to help people avoid the biases they seem to so naturally adopt.

References

1. Baeza-Yates, R., Saint-Jean, F., Castillo, C. Web dynamics, age and page quality. In: *Proceedings of SPIRE* (2002).

2. Berners-Lee, T., Cailliau, R., Groff, J., and Pollermann, B. World Wide Web: the information universe. *Electronic Networking: Research, Applications, and Policy 2(1), 52-58* (1992).

3. Brin, S., Page, L. The anatomy of a large-scale hypertextual Web search engine. In: *Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia 107-117* (1998).

4. Cho, J., Adams, R. E. Page quality: In search of an unbiased web ranking. Technical report, UCLA Computer Science Department (2003).

5. Cho, J., Roy, S. Impact of search engines on page popularity. In: *Proceedings of the Thirteenth International World Wide Web Conference* (2004).

6.Church, K., Keane, M.T., Smyth, B. The first click is the deepest: assessing information scent predictions for a personalized search engine. In: *Proceedings of the 3rd Workshop on Empirical Evaluation of Adaptive Systems. 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (2004)

7. Freyne, J., Smyth, B. Collaborative search: a live user trial. In: *Proceedings of the 26th European Conference on IR Research, ECIR.04, Sunderland, UK* (2004)

8. Hölscher, C., and Strube, G. Web search behavior of internet experts and newbies. *Proceedings of the 9th International WWW conference. Amsterdam, The Netherlands, 337-346* (2000).

9. Joachims, T., Granka, L., Pang, B., Hembrooke, H. Gay, G. Accurately interpreting clickthrough data as implicit feedback. In: *Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR)* (2005).

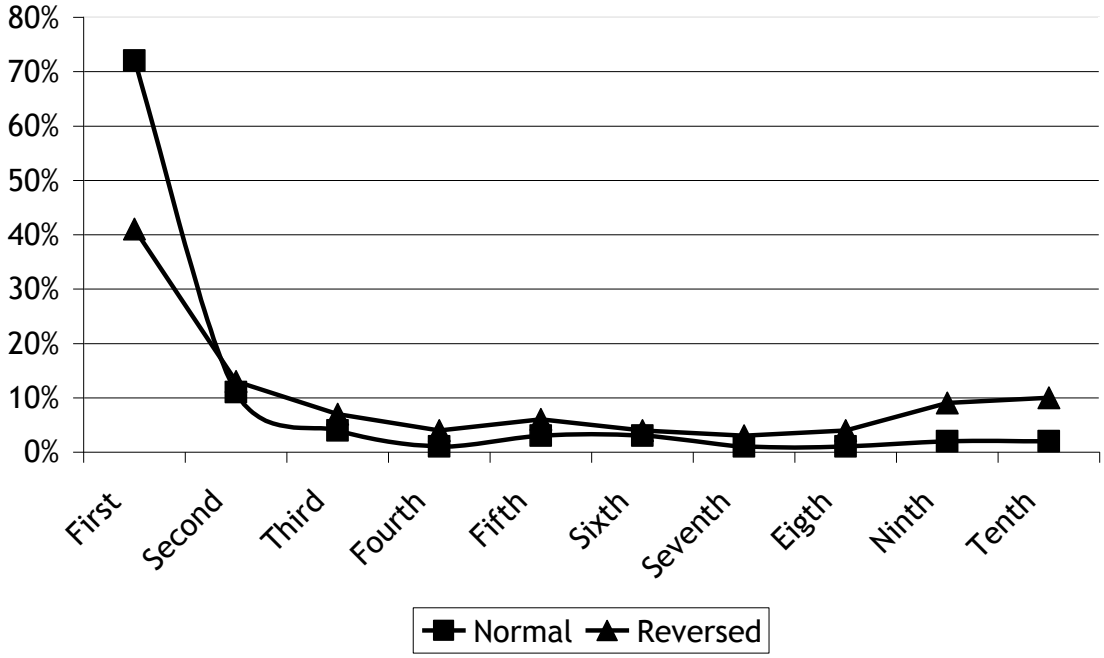

Figure 1: The normalised percentage of first-clicks by position in the list in the normal & reversed conditions