Fading Away: Dilution and User Behaviour

Paul Thomas
CSIRO ICT Centre
paul.thomas@csiro.au

Falk Scholer
School of Computer Science
and Information Technology
RMIT University
falk.scholer@rmit.edu.au

Alistair Moffat
Department of Computing and
Information Systems
The University of Melbourne
ammoffat@unimelb.edu.au

ABSTRACT

When faced with a poor set of document summaries on the first page of returned search results, a user may respond in various ways: by proceeding on to the next page of results; by entering another query; by switching to another service; or by abandoning their search. We analyse this aspect of searcher behaviour using a commercial search system, comparing a deliberately degraded system to the original one. Our results demonstrate that searchers naturally avoid selecting poor results as answers given the degraded system; however, the depth of the ranking that they view, their query reformulation rate, and the amount of time required to complete search tasks, are all remarkably unchanged.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*performance evaluation*.

General Terms

Experimentation, measurement.

Keywords

Retrieval experiment, evaluation, system measurement.

1. INTRODUCTION

While carrying out a search, users have a number of tactics available to them. Intuitively, it seems likely that these tactics or behaviours will vary based on the quality of the results that are returned by the retrieval system. For example, other things being equal, a user who cannot find any relevant items on the first page of search results might be more inclined to reformulate their query (by entering another query into the search interface) than a user who has found a large number of relevant items. Possible tactics when using an apparently ineffective system include:

1. Looking further in the results list, visiting pages beyond the first, hoping that the results improve;
2. Submitting another query, hoping for better results;
3. Switching to a different search engine and entering the same query, hoping that it provides better results;
4. Trying to find the information through other techniques, for example by browsing.

We investigate the first two possibilities, reporting on differences in user behaviour when a standard retrieval system is compared to an adjusted system in which results are diluted by inserting non-relevant answers. Our results indicate that searchers remained attentive to the task in the degraded system, and adapted their behaviour to avoid clicking on non-relevant snippets. However, all other aspects of their behaviour were remarkably consistent, including the amount of time spent on tasks; the number of query reformulations undertaken; and their perceptions of search difficulty.

2. METHODS

We designed a user experiment to explore ways in which behaviour changes with retrieval quality. A total of $n = 34$ participants, comprising staff and students from the Australian National University, carried out six search tasks of differing complexity, covering the *remember*, *analyse* and *understand* tasks of Wu et al. [7] but modified for our context. On commencing a task, users were shown a result page for an initial "starter" query that was constant across users. They were then free to explore the results list, including being able to open documents, to view further results pages, and to enter follow-up queries. Once any document was opened for viewing, participants were asked to indicate whether or not it was relevant to their search task, before returning to the search results listing. The search interface prevented tabbed browsing, and while a document was being viewed it replaced the results page. Participants were not given an explicit time limit for any task, but were told they could move on when they felt ready.

The search results displayed to participants were sourced from the Yahoo! API, and presented in the usual way as an ordered list consisting of query-biased summaries, with ten results per page. No branding from the underlying search service was shown. Without telling our participants, we simulated search systems of two different effectiveness levels by showing results in one of two modes: *full*, where the ranking obtained from the search service was displayed in its original form; and *diluted*, where the original results were interleaved with answers from a related but incorrect query [5]. Dilution was operationalised by leveraging the capacity-enhancing (and obfuscatory) power of "management-speak": the original stakeholder information need was actioned going forward by enhancing it through the win-win inclusion of a jargon competency chosen randomly from a list of outside-the-box strategies, thereby disempowering the results paradigm. For example, if the task was to "find

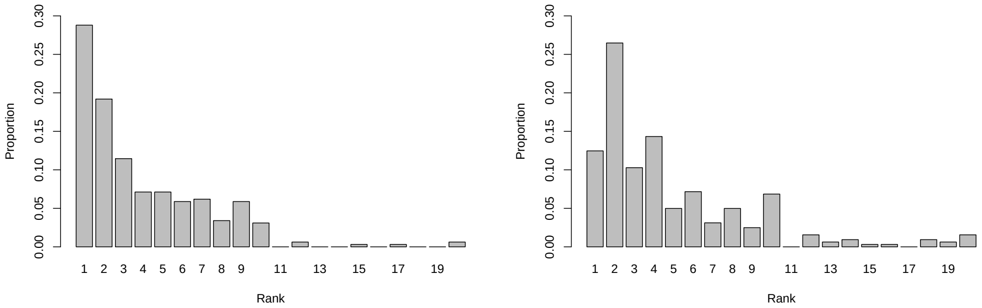

Figure 1: Normalised total click positions across participants and tasks, for *full* queries (left) and *diluted* queries (right).

the Eurovision Song Contest home page", a user's initial *full* query might be "*eurovision*"; whereas in the *diluted* system half of the results displayed might instead be derived from the query "*eurovision best practice*". There were a small number of queries issued for which it was not possible to generate five such results; these 22 out of 5930 page interactions are excluded from the analysis below.

Most interactions with the search system were logged while participants carried out the six search tasks, including: submitted search queries; clicks on snippets in order to open documents for viewing; assessments of document usefulness; and the point of gaze on the screen, captured using an eye tracker. Task order was balanced across the participants and topics so as to minimise the risk of bias; similarly, whether the *full* or *diluted* approach got applied for each participant-task combination was pre-determined as part of the experimental design.

3. RESULTS

User behaviour, and the differences caused by the *full* and *diluted* query treatments, can be measured in a range of ways.

User click behaviour: The normalised click frequency at each rank position in the answer pages is shown in Figure 1. In the *diluted* retrieval system the "incorrect but plausible" documents were inserted in positions $1,3,5,7$ and 9. The pattern of click behaviour demonstrates that our experimental manipulation was successful: for the *full* search results, the click distribution follows the expected pattern of users clicking more frequently on items that are higher in the ranked list [1], whereas users of the *diluted* system were less likely to click answer items in the odd positions. Note that position bias – the propensity for searchers to select items that occur higher in a ranking, possibly because they "trust" the underlying search system [3] – exists in both systems. In particular, all of the odd-numbered rank positions in the *diluted* system are equally "bad", but participants still favoured items higher in the ranking.

A second check to confirm that our system dilution had an impact on search effectiveness is to consider the rates at which users saved documents that they viewed (that is, the likelihood that a document was found to be relevant after it was clicked). The mean rate is 0.733 for the *full* system, compared to 0.597 for the *diluted* system, a statistically significant difference (t-test, $p < 0.05$).

While Figure 1 establishes that our user study participants responded differently in terms of rank-specific click behaviour, the high-level aggregated click behaviour across all participants and search tasks was not distinctive: in total (all tasks, and all users) there were 323 clicks for the *full* system, and 322 for the *diluted* system. Unsurprisingly this difference is not statistically significant (χ^2 test, $p = 0.97$). The number of items that were determined as being useful was also similar in the two conditions: 201 for *full*, and 214 for *diluted* (χ^2 test, $p = 0.52$). Our participants needed to read a remarkably similar number of documents, and a remarkably similar number of useful documents, to satisfy the (assigned) needs regardless of the search system.

	1st results page	2nd results page
full	207	15
diluted	212	22

Table 1: Total page views, summed across users and topics, for the *full* and *diluted* retrieval systems.

Given this difference in click rates, it is reasonable to expect other changes in behaviour and we consider this below.

Depth of result page viewing: When presented with a search results page, the user chooses which snippets require further evaluation. In line with commercial search engines, our experimental participants were presented with ten answers per page, with the option of accessing subsequent results pages.

Faced with a relatively poor quality results list, a plausible strategy for a user who is looking for an answer document is to look further down the results page. Table 1 shows the frequency with which results pages were viewed (that is, the user visited a results page and looked at one or more items on the screen as recorded using eye-tracking), summed across users and queries. When using the *full* system, participants moved on to the second page of results for 15 out of 207 issued queries (with a corresponding mean page depth of 1.07), while in the *diluted* system the second results page was visited for 22 out of the total of 212 queries that were issued (a mean page depth of 1.10). The difference in depth was not significant (χ^2 test, $p = 0.34$). No participants viewed results beyond the the second page with either system.

Figures 2 and 3 provide a more detailed view of gaze behaviour, showing the deepest rank position that searchers examined while carrying out a query, and the last rank position that was viewed before finishing the query. The distributions of the lowest rank positions viewed are similar between the *full* and *diluted* systems: both show peaks at rank positions 7 (the last item above the fold) and 10 (the last item in each page of search results). The distribution of the last position viewed before finishing a query (which arises when either enough relevant items have been found, or the user types a fresh query) are also broadly similar. However, for the *diluted* system, rank position 1 has a larger proportion of the probability

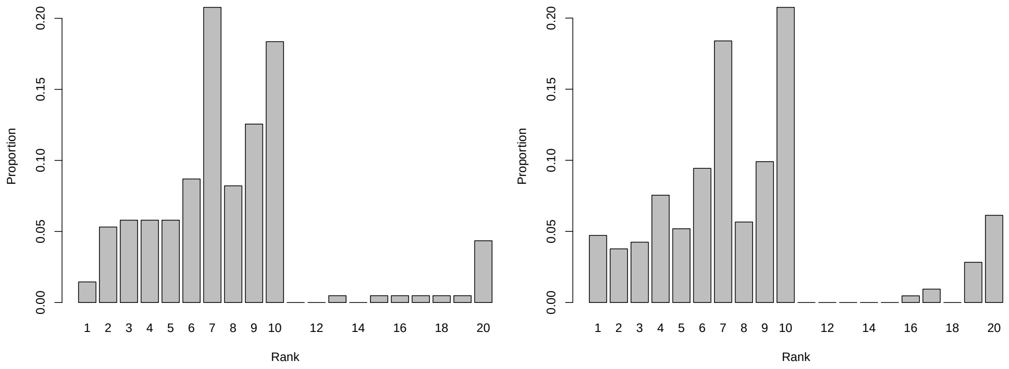

Figure 2: Deepest rank position viewed, averaged across topics and participants, for *full* queries (left) and *diluted* queries (right).

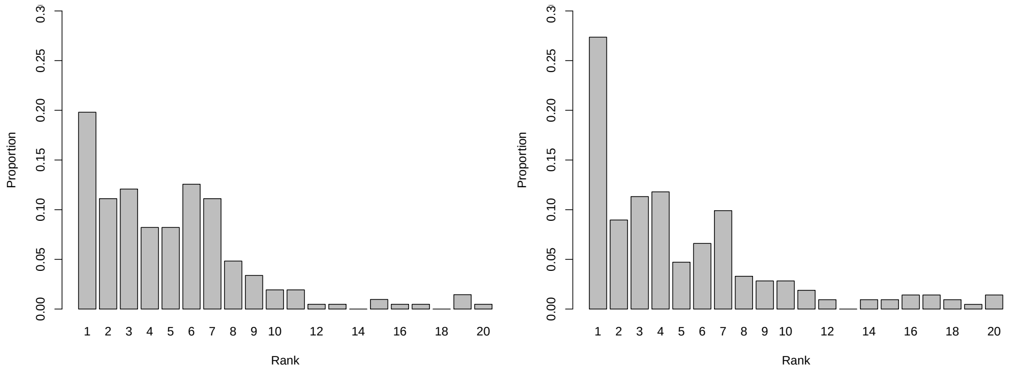

Figure 3: Final rank position viewed, averaged across topics and participants, for *full* queries (left) and *diluted* queries (right).

mass. A possible reason is that searchers mentally compare answers as they view items in the results list, and most users scan at least the top few items. The *diluted* system is likely to have a non-relevant document in position one, and so reviewing that snippet may serve as a final confirmation, before the user commits to a click on a deeper-ranked snippet from the underlying *full* results.

Query reformulation: A second way in which a user might respond to search systems of differing quality is to change the rate at which they stop looking through the current set of search results, and instead enter a new query.

The number of queries used by participants when carrying out their search tasks is shown in Figure 4. Overall the number was low for both systems, with a median of 1 and 2 queries (0 and 1 reformulations) for the *full* and *diluted* results, respectively. This difference was not statistically significant (Wilcoxon signed-rank test, $p = 0.46$).

Ability to identify relevant answers: When a retrieval system serves unhelpful answers, it might be that the ability of the searcher to identify useful answers is similarly affected. However, based on our experiments, the mean rate at which clicked items were saved as being relevant was 0.787 for the *full* system and 0.747 for the *diluted* system, showing no significant difference (t-test, $p = 0.25$). Thus the ability of users to identify relevant answers, once documents have been selected for viewing via their snippets, did not differ between the experimental treatments.

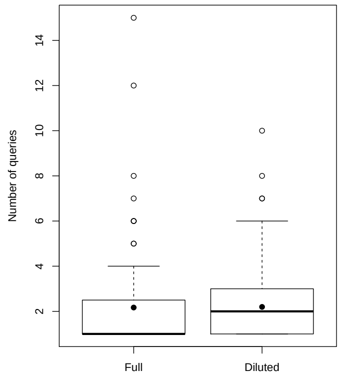

Figure 4: Number of queries per task, for *full* and *diluted* queries.

Time spent on tasks: While depth of viewing and query re-formulation do not show significant differences in searcher behaviour, it could still be the case that using an inferior system makes querying slower. Differences in system quality might alter the time spent by users when viewing and processing result pages. However, the average gaze duration when viewing snippets, measured as the sum of fixation durations that occurred in the screen area defined by each search result summary, was 0.586 second for *full* queries and 0.589 seconds for *diluted* queries. This difference was not statistically significant (t-test, $p = 0.89$).

Differences could also occur at a higher level of system interaction. The mean time that participants spent working on each search task, including viewing search result pages, viewing selected documents, and making relevance decisions, was 2.70 minutes for the *full* treatment, and 2.54 minutes for the *diluted* one. This difference was not statistically significant (t-test, $p = 0.62$).

Finally, we consider the interaction between time and query reformulations. When using the *full* system, participants entered an average of 1.50 queries per minute while completing each task. For the *diluted* system, the rate was 1.52 queries per minute. The difference was not significant (t-test, $p = 0.95$).

Overall, these results indicate that the quality of the search system did not affect the rate at which participants were able to process information on search results pages, or how much time they spent working on tasks before feeling that they had achieved their goals. The only significant difference between the two treatments was the click distribution, and the rate at which clicked documents were judged to be useful.

Searcher assessment of task difficulty: After carrying out each search task, experimental participants were asked to answer two questions: "How difficult was it to find useful information on this topic?", and "How satisfied were you with the overall quality of your search experience?". The 5-point response scale for these questions was anchored with the labels "Not at all" (assigned a value of 1) and "Extremely" (assigned a value of 5).

Searchers found the tasks relatively easy to complete: the median response rate for the search difficulty question was 2 for both the *diluted* and *full* systems; this difference was not significant (Wilcoxon test, $p = 0.73$). Satisfaction levels were also highly consistent between the two systems, with a median response level of 4 for both systems (Wilcoxon test, $p = 0.91$). Overall, there were no systematic differences in participants' perceptions of search difficulty or the overall experience resulting from the two different treatments.

4. DISCUSSION AND CONCLUSIONS

It seems "obvious" that user behaviour will be influenced by the quality of results that returned by a search service. Seeing many poor results near the start of an answer list may influence the user's decision about whether to continue viewing subsequent answer pages, to enter a new query, or to abandon the search altogether. Previous work has supported this view. For example, in a study of 36 users completing 12 search tasks with different search systems, Smith and Kantor [4] found that users adapted their behaviour: when given a consistently degraded search system, they entered more queries per minute than users of a standard system; similarly, a higher detection rate (the ability to identify relevant answers) was observed for users of degraded systems.

However our study, in which 34 subjects carried out search tasks using an evenly balanced combination of *full* and *diluted* search systems, contrasts strongly with that intuition and previous findings. Overall, searchers took around the same amount of time to complete their tasks in both experimental treatments; were able to save a similar number of documents as being relevant; exhibited consistent viewing behaviour when looking at the search results lists returned by the treatments; and did not perceive significant differences in the difficulty of carrying out tasks with both systems. The key difference in participant behaviour was their click rate at particular ranks: in essence, they successfully avoided poor answers, as demonstrated by the shift in the click probability mass, shown in Figure 1.

A possible explanation for the divergence in observed user behaviour between the two studies may be the context in which the searches were carried out. Participants in the Smith and Kantor study were instructed to *"find good information sources"* for an unspecified *"boss"*, with an incentive to *find the most good and fewest bad sources possible* [4]; participants were not constrained in the amount of time that they could spend on a task. In contrast, our subjects were instructed that they would complete *a sequence of . . . web search tasks* and were advised to *spend what feels to be an appropriate amount of time on each task, until you have collected a set of answer pages that in your opinion allow the information need to be appropriately met*. The overall expectations were therefore different: in the Smith and Kantor study, participants were given the goal of maximising relevance by finding as many good answers as possible; in our study, participants were "satisficing", having been requested to decide for themselves when an appropriate number of answers had been found.

Alternatively, it may be that our *diluted* system, while certainly poorer in overall quality (in the sense that non-relevant answers were introduced into the ranking), was not poor enough to induce different behaviour. Smith and Kantor used results typically from the 300th position in Google's results: even today, these are unreliable for the simplest of our topics, and in 2008 will almost certainly have produced a poor result set. Importantly, our *diluted* system always included a few high-ranked results.

Either way, our results raise an important question about how the effectiveness of search systems should be analysed. While some fine-grained aspects of user clicking behaviour differed between the *full* and *diluted* treatments, the majority of behaviours did not. This outcome is in line with previous results that found little relationship between user behaviour and system quality as measured by common IR evaluation metrics such as MAP [6]. The question then becomes one of whether even a significant improvement in effectiveness, as measured by some metric, actually results in improved task performance. In future work, we therefore plan to systematically investigate different levels of answer-page dilution, to establish guidelines for the extent of practical differences that need to be present in search systems for measurable disparities in user behaviour to manifest. We also plan to explore the issue of the impact that specific variations in task instructions have on searcher behaviour through a controlled user study in a work task-based framework [2].

References

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proc. SIGIR*, pages 3–10, Seattle, WA, 2006.

[2] P. Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1):71–90, 2000.

[3] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. SIGIR*, pages 154–161, Salvador, Brazil, 2005.

[4] C. Smith and P. Kantor. User adaptation: good results from poor systems. In *Proc. SIGIR*, pages 147–154, Singapore, 2008.

[5] P. Thomas, T. Jones, and D. Hawking. What deliberately degrading search quality tells us about discount functions. In *Proc. SIGIR*, pages 1107–1108, Beijing, China, 2011.

[6] A. Turpin and F. Scholer. User performance versus precision measures for simple web search tasks. In *Proc. SIGIR*, pages 11–18, Seattle, WA, 2006.

[7] W.-C. Wu, D. Kelly, A. Edwards, and J. Arguello. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proc. 4th Information Interaction in Context Symp.*, pages 254–257, Nijmegen, The Netherlands, 2012.