Eye-Tracking Reveals the Personal Styles for Search Result Evaluation

Anne Aula, Päivi Majaranta, and Kari-Jouko Räihä

Tampere Unit for Computer-Human Interaction (TAUCHI),
Department of Computer Sciences, FIN-33014 University of Tampere, Finland
{aula, curly, kjr}@cs.uta.fi

Abstract. We used eye-tracking to study 28 users when they evaluated result lists produced by web search engines. Based on their different evaluation styles, the users were divided into *economic* and *exhaustive* evaluators. Economic evaluators made their decision about the next action (*e.g.*, query re-formulation, following a link) faster and based on less information than exhaustive evaluators. The economic evaluation style was especially beneficial when most of the results in the result page were relevant. In these tasks, the task times were significantly shorter for economic than for exhaustive evaluators. The results suggested that economic evaluators were more experienced with computers than exhaustive evaluators. Thus, the result evaluation style seems to evolve towards a more economic style as the users gain more experience.

1 Introduction and Related Work

To date, the process of evaluating result lists of web search engines has received little attention. Eye-tracking is a promising method for this purpose as it provides information on visual information processing. In HCI, eye-tracking has been used to study the usability of web pages [3], menu searching [1, 2], and information searching from web pages [4] and hierarchical displays [6], among others. Three previous eye-tracking studies have specifically focused on search result evaluation.

Salojärvi et al. [9] aimed at inferring the relevance of newspaper headings from the features of eye-tracking data. The features they found useful for the relevancy prediction were, for example, fixation count, total fixation duration, and pupil size.

Granka et al. [5] studied how users browse result listings and how they select links. In their study, the participants entered their own queries for 10 tasks. Their results suggested that users spend most of the time fixating on the first and the second result before their initial click. The third and following results get significantly less fixation time. Their results also indicated that users tend to follow a sequential strategy in scanning the results by going from top to bottom until they follow a link.

In Klöckner et al. [7], all participants saw the same result page with 25 results. Their task was to collect information from this list for one open-ended search task. A majority of the users (65%) used a *depth-first strategy* where only the results above the selected link are evaluated before the selection. 15% of the users used a *breadth-first strategy* where they looked through all the results before opening any documents. The remaining 20% of the participants used a mixture of these strategies.

M.F. Costabile and F. Paternò (Eds.): INTERACT 2005, LNCS 3585, pp. 1058 – 1061, 2005.

Our aim was to deepen the understanding of the results evaluation process with a semi-controlled study. Our participants first saw a pre-defined search result page for each task making it possible to compare their gaze data when viewing exactly the same stimulus. However, we aimed for as realistic gaze behaviour as possible. Thus, immediately after seeing the pre-defined page, the participants could modify the query, enter a URL, or select a result from the pre-defined page. This semi-controlled method enabled us to find the differences in the participants' evaluation styles in a realistic search situation while also preserving the control over the first stimulus.

2 Methods

All of the pre-defined queries were submitted to Google and the corresponding result pages were saved as local files. The 10 queries were chosen so that three of them were poor (no relevant results in the list), three were good (more than five relevant results), and the remaining four were mixed. We used Tobii 1750 remote eye-tracker with its default 17 inch TFT monitor with a resolution of 1280 × 1024. 42 students from different majors participated. Due to technical problems, data from 4 participants was excluded. This paper reports the preliminary results from the first 28 participants (11 females 17 males; average age 23.7 years), who were all experienced in using Google.

The participants were told that the purpose was to study their normal information searching with search engines, as well as to test the eye-tracker for pupil size measurements during web use. The cover story ensured that the participants did not concentrate on their eye-movements. The eye-tracker's calibration was tested before each task and it was re-calibrated if needed. The tasks were presented on slips of paper in random order. The participant proceeded to the task-specific result listing by selecting it from a list and then continued the searching normally. In the end, the participants filled in a background questionnaire. Finally, they were debriefed.

3 Data Analysis and Results

The results are based on the gaze data in the pre-defined result page until the user selected a result, formulated a new query, or exited the page otherwise. For the analysis, we defined Areas of Interest (AOI) in ClearView eye-gaze analysis software provided by Tobii. The area above the result listing formed one AOI, as did each individual result. We then collected all successive fixations to each AOI and calculated their cumulative fixation duration.

For analysing the evaluation styles, we developed a static visualization that presents the order in which each AOI is visited [8]. To enable comparison of the evaluation styles of different individuals and the effects of good and poor result lists, all the visualizations (280 visualizations in total) were printed out on paper and visually inspected by each author, first separately and then in face-to-face meetings. The goal of the inspection was to find patterns in the evaluation styles and to group the visualizations accordingly. As a result of this analysis, two groups of users with different result evaluation styles were identified. We call these groups *economic* (46% of the participants) and *exhaustive evaluators* (54% of the participants).

In the pre-defined result lists, there were six or seven results visible in the result list without scrolling. In over 50% of the tasks, economic evaluators scanned at most half (three) of those results before their first action. Exhaustive evaluators, on the other hand, evaluated in most of the tasks more than half of the visible results or even scrolled the results page to view all of the results before performing the first action. Examples of economic and exhaustive evaluation strategies are presented in Fig 1.

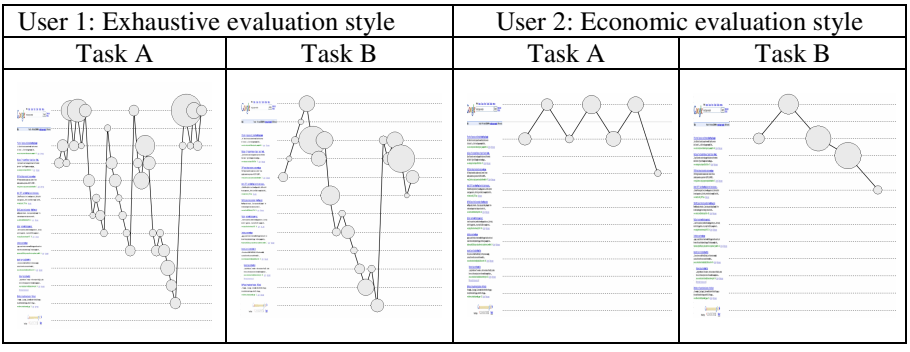

Fig. 1. Examples of evaluation styles. The *y* axis shows the vertical position in the search result page with a compact representation of the result page shown on the left side. The *x* axis shows the order in which different AOIs were visited. The size of the circle corresponds to the time spent on each AOI, the largest circles are approximately 3 seconds. In task A, the results were irrelevant to the task (both users chose to reformulate the query) and in task B, most of them were relevant (both users followed the second link).

An independent samples t-test showed that the time before the first action was significantly shorter for economic than for exhaustive evaluators, $t(26) = -3.1$, $p < .01$. In computer experience (measured by multiplying the frequency of computer use and the years of usage), a marginal difference between the two groups was found with economic being more experienced than exhaustive evaluators, $t(26) = -1.9$, $p = .07$. There was also a correlation between the average fixation duration and the evaluation style: the more economic the searcher was, the shorter were the fixations (Pearson's two-tailed $r = -.44$, $p < .05$).

In the result pages differing in quality, we found that when the query was good and most of the results were relevant, the economic evaluation style was especially beneficial. In this case, the task times of economic evaluators were shorter than those of exhaustive evaluators (average times 74.9 and 109.0 seconds), $t(26) = -2.0$, $p = .05$. In tasks with poor results, the task times did not differ.

4 Discussion

Our results suggested that economic evaluation style was more efficient in search tasks, especially when the quality of the results was good. Thus, it seems that it is beneficial to quickly click on the relevant-looking result instead of carefully trying to choose the best one. The results also identified a large group of exhaustive evaluators.

They, possibly due to their lack of expertise, carefully evaluate the results before following a link or re-formulating a query. Thus, they are more dependent on the result summaries given by the search engine.

Economic and exhaustive evaluation styles resemble the depth-first and breadth-first strategies by Klöckner et al. [7]. However, their participants scanned a list of 25 results with a task of selecting relevant results from the list. In contrast, our participants could go through the pre-defined result list, select results, or modify the query according to their will. This setup presumably enabled them to employ their normal scanning styles. Therefore our results cannot be directly compared with those by Klöckner et al. However, our results indicated a large group of less-experienced users who evaluate results below the one that gets selected and even thoroughly scan irrelevant results. Granka et al. [5] suggested that users tend to scan only the results above the selected link. As they did not report the computer experience of the participants, it is possible that their participants were experienced and thus, employed economic strategies as suggested by our results.

The data analysis is still ongoing. We are analyzing the data from the rest of the participants and in other pages than the pre-defined ones. We are also developing metrics for analyzing the differences in the gaze paths of the users with different evaluation styles. In the quest for *proactive search interfaces* [9], we will analyze the data in order to infer the relevance of the individual results from the gaze data.

References

1. Aaltonen, A., Hyrskykari, A., and Räihä, K.-J.: 101 spots, or how do users read menus? In Proc. CHI 1998, ACM Press (1998) 132-139
2. Byrne, M.D., Anderson, J.R., Douglass, S., and Matessa, M.: Eye tracking the visual search of click-down menus. In Proc. CHI 1999, ACM Press (1999) 402-409
3. Ellis, S., Cadera, R., Misner, J., Craig, C.S., and Lankford, C.P.: Windows to the soul? What eye movements tell us about software usability. In Proc. 7th Annual Conf. Usability Professionals Association, Washington D.C. (1998)
4. Goldberg, J.H., Stimson, M.J., Lewenstein, M., Scott, N., and Wichansky, A.M.: Eye tracking in web search tasks: Design implications. In Proc. ETRA'02, ACM Press (2002) 51-58
5. Granka, L., Joachims, T., and Gay, vcG.: Eye-tracking analysis of user behavior in WWW search. In Proc. SIGIR'04, ACM Press (2004) 478-479
6. Hornof, A.J., and Halverson, T.: Cognitive strategies and eye movements for searching hierarchical computer displays. In Proc. CHI 2003, ACM Press (2003) 249-256
7. Klöckner, K., Wirschum, N., and Jameson, A.: Depth- and breadth-first processing of search result lists. In Proc. CHI 2004, ACM Press (2004) 1539
8. Räihä, K.-J., Aula, A., Majaranta, P., Rantala, H., and Koivunen, K.: Static visualization of temporal eye-tracking data. In Proc. INTERACT 2005, Rome, September 2005
9. Salojärvi, J., Kojo, I., Jaana, S., and Kaski, S. Can relevance be inferred from eye movements in information retrieval? In Proc. WSOM'03 (2003) 261-266