USING EYE TRACKING TO EVALUATE ALTERNATIVE SEARCH RESULTS INTERFACES

Rachana S. Rele and Andrew T. Duchowski
Clemson University
Clemson, SC 29634

Surveys have shown that 75% of users get frustrated with search engines and only 21% find relevant information. Inability to find relevant results can be partially attributed to cluttered results pages and failure in constructing Boolean queries. This research used sixteen subjects to evaluate two types of search results interfaces using four tasks while measuring performance and studying their ocular behavior using a Tobii 1750 eye-tracker. The two interfaces used were list interface, commonly seen on many search engines and a tabular interface presenting information in discrete categories or elements of the result's abstract. Quantitative comparisons of two interfaces are made on performance metrics such as time and errors, process metrics such as fixation durations, number of fixations, and eye movement transitions from one element or category of the abstract. Subjective data was collected through post-task and post-test questionnaires. The results did not show any significant difference in performance between the two interfaces, however, eye movements analysis provide some insights into importance of search result's abstract elements such as title, summary, and URL of the interface while searching.

INTRODUCTION

According to PEW/ American Life Project (Fallow, 2005) web searching is the second most popular activity on-line, first being email. Web users spend 70% or more of their time searching on the web (RealNames, 2000). In a "Search Rage" study conducted by WebTop (2000) 75% of the respondents reported significant amount of frustration of some significant degree and 86% of the users said that searching could be more efficient. In a more recent study (Käki, 2005), it was found that users found a relevant result in only 40% of the first selections they made on list format of search results. Design of search query input and search results interfaces can positively contribute towards finding relevant information.

Query Input Interface

With the aim of improving users search efficiency at the query input interface level, Bandos and Resnick (2004) introduced examples of Boolean operators on the search query interface. Another problem with search input interface is that of employing Boolean logic in search queries, which can potentially increase the precision of search. However, it can be conjectured that users will require additional time to formulate the query which will cause an increase in the overall search time. Since, search time is considered to be the most essential determinant in evaluating search performance by users (Drori, 2003), additional user behavior research needs to elicit whether users would prefer to spend effort or time in formulating perfect queries or finding information on a structured interface.

Search Results Interface

The most commonly used search results interface by commercial web search engines is a list format of results. Resnick et al. (2001) designed a tabular interface in which columns of the table corresponded to the different elements of the abstract presented in the list interface. The tabular interface in Resnick et al.'s study supported faster scanning of results in comparison with the list interface. The subjective data showed that for the tabular layout, users scanned only one field for all options until they found one that met their match criterion, and this layout was also the preferred among the two interfaces. Similar research (Dumais and Chen, 2001; Käki, 2005, Drori, 2003) has developed alternatives to this interface in which results are classified into clusters of information. These studies have shown to improve search performance in comparison to the list format of presenting results. Furthermore, this increase in performance can be attributed to information chunking in the form of categories which could have allowed users to prioritize their attention.

Granka et al. (2004) studied user's ocular behavior on list interface of search results in which they observed the amount of attention each abstract on the list interface received and the corresponding clicks on these abstracts. The findings from this study suggest that users devote lesser attention to abstracts located below the page break. This behavior may cause reduction in precision of search if relevant results are located below the page break. Hence, list interface may prove inefficient when the search engine does not provide the most sought for results on the first results page between the ranks 1 through 5.

Klockner et al. (2004) analyzed eye movements over a page containing 25 results listed in the format used by the Google search engine. More than half of the users studied, 65 % applied a strategy in which the user examined each entry in the list in turn, starting from the top, deciding whether to open it. Fifteen percent adopted a strategy in which they looked ahead at a number of results in the list, revisiting and opening only the most promising ones. The remaining 20% showed a

mixed strategy, looking ahead at the next few entries before deciding which documents to open.

Salvogarvi et al. (2003) studied the amount of pupil dilation on the list of search results using eye tracking technology. They found that pupil dilation increases while viewing relevant abstracts. However, this study used three subjects making it difficult to generalize results across diverse user population.

The primary hypothesis of this research is that tabular interface will increase efficiency and accuracy in scanning search result which will be achieved through spatial grouping of results into distinct element/category columns.

METHODOLOGY

Participants

Sixteen participants (6 F, 10 M) with ages ranging from 20 to 29 years (Mean = 24.5) were recruited for this study. All the users had a minimum of 5 years internet experience, and searching information was one of their daily internet activities, with Google being their primary search engine.

Stimulus

List interface had the same look and feel of the list interface used by the Google search engine. However, sponsored advertisements were removed, eliminating any effects due to these links. Figure 1 shows the list interface used for this study.

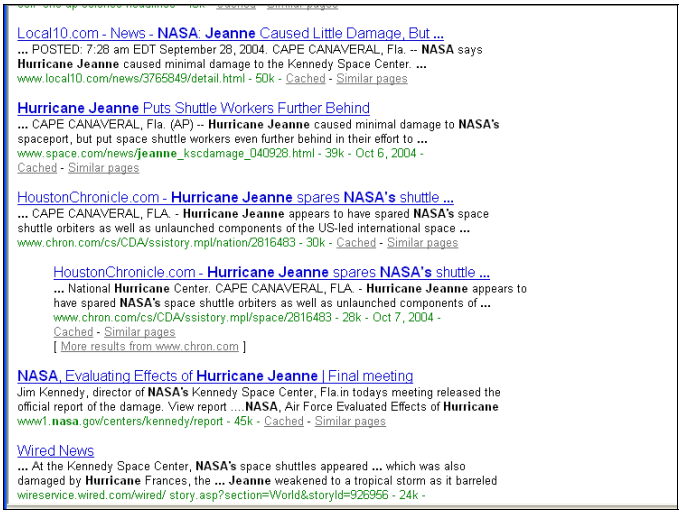

Figure 1. List interface derived from Google

A tabular interface was designed such that the every element in the list interface had a corresponding column in the tabular interface. The columns were arranged starting from left such that the elements of information appeared in the same sequence as that on the list interface (Figure 2).

Log analysis studies (Silverstein, 1998; Jansen et al., 2000) showed that 90% users view only the first page of search results. With respect to this, the current research evaluated interface layouts based on the viewing behavior, and performance on the first ten search results or the first results page.

Figure 2. Tabular interface

Tasks

Two informational and two navigational search tasks were designed based on web search task taxonomy (Broder, 2002). Informational tasks are one in which information is assumed to be present on one or more web pages. Navigational tasks are designed to arrive at a particular website or URL. Keywords for the tasks were pre-decided by the experimenters to make the results comparable across participants. These keywords were used to retrieve results from the Google search engine for the four tasks. A search engine was simulated using the first page retrieved for each of the four queries. Following are examples of an informational and a navigational query used in this study.

- Find information on the report published by NASA, detailing Hurricane Jeanne's damage to its space center at Florida. (Informational query)
- Find the home page of Michael Murray, a mathematician. (Navigational query)

Experimental Design

A two-factor factorial design was used with two factors being interface type (List and Tabular) and task type (Informational and Navigational) each at two levels. Eight task-interface treatment combinations were obtained, out of which four were assigned to each participant. The sequence of tasks was counterbalanced across subjects.

Apparatus & Settings

The study used a PC integrated with a Tobii 1750 binocular eye tracker with 17" display having a maximum resolution of 1280 X 1024 pixels. The eye tracker has a tracking rate or the frame rate of 50 Hz, and looks like a

normal computer display with cameras and illuminators hidden behind filters. Hence, the eye tracking is nearly invisible to the user. The Tobii eye-tracker is a bright-pupil eye tracker that uses a camera with a high resolution and large field of view to capture images of the subject's eyes. NIRLED's (Near Infra Red Light Emitting Diodes) are used to generate even lighting and reflection patterns off the subject's eyes. The Tobii screen subtends a visual angle of 28 degrees horizontally and vertically at the participants' eyes from a distance of 60 centimeters. Since the user tasks involve scanning as well as reading on the interface, the fixation duration is chosen to be 40ms and the fixation size is chosen to be 20 pixels. Figure 3 shows the Tobii eye tracker used for this experiment.

Figure 3. PC integrated Tobii 1750 binocular eye-tracker

Procedure

Participants were screened for a minimum of 5 years web experience. The eye-tracker was calibrated to the participant's eyes using a 16-point calibration. Participants were then familiarized (un-paced) with the tabular interface and were then given instructions for the study. At the beginning of each task a query interface appeared with the keywords used for the tasks in the search field. Users were required to click on the "Search" button to retrieve results for the query. There was only one right answer to the query, and hence clicking on a wrong link resulted in a page saying "To continue search, go back to the search results". The trial was terminated when participant found the correct result or if he or she chose to terminate the search session. A post-task questionnaire was administered to gather data specific to a tasks and a final post-test questionnaire to find overall impressions about the two interfaces. Sixteen participants performed the four tasks, two on list interface and two on the tabular interface.

Dependent Variables

Performance measures.

Search time

Number of wrong result choices

Process measures.

Mean fixation duration

Number of fixations on different categories such as Title, Summary, and URL

Probability of making transitions from one category of an abstract (e.g., from Title category of an abstract) to the same category (e.g., to Title category of another abstract) in the next abstract

Subjective measures.

Perceived time of task completion

Perceived accuracy in choosing results

Preference of one interface over the other

Overall satisfaction with the interface

RESULTS

Performance measures

There was no interaction found between the type of task and the type of interface for the performance measures of time and number of wrong results choices. There was no significant difference in the search time ($F(1, 60) = 2.34$, $p>0.05$) and number of wrong result choices ($F(1, 60) = 0.16$, $p>0.05$) on the list and tabular interfaces. Similarly, time and number of wrong results choices did not significantly differ based on the type of task with $F(1, 60) = 0.77$, $p>0.05$ and $F(1, 60) = 0.03$, $p>0.05$, respectively.

Process measures

The mean fixation durations on the two interfaces did not significantly differ ($F(1, 60) = 1.99$, $p>0.05$). The type of tasks revealed no significant difference ($F(1, 60) = 0.25$, $p>0.05$) in the mean fixation duration. Figure 4 shows a comparison between the mean fixation durations on the two interfaces for navigational and informational tasks. Figures 5 and 6 indicate the fixation durations by the intensity of colors on the list and tabular interface respectively.

The number of fixations in the Title category of the two interfaces did not significantly differ ($F(1, 60) = 0.55$, $p>0.05$). However, the number of fixations in the Summary category significantly differed ($F(1, 60) = 7.4$, $p=0.008$) for the two types of tasks, navigation tasks requiring more number of fixations (42.68) than the information task (24.15). There was a significant difference ($F(1, 60) = 11.55$, $p=0.001$) in the numbers of fixations falling in the URL category for list interface and the tabular interface. Figure 7 shows a comparison between the list and tabular interfaces in terms of number of fixations in three different categories.

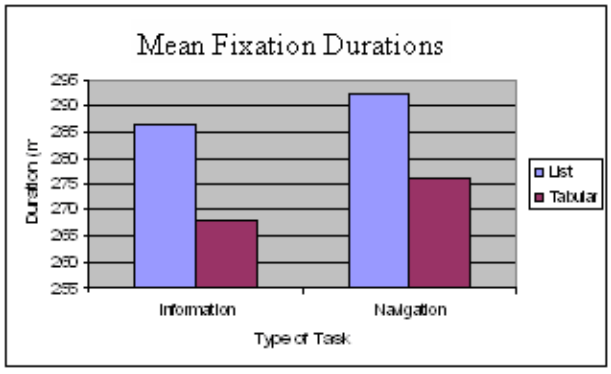

Figure 4: Mean fixation durations on the two interfaces, for two types of tasks

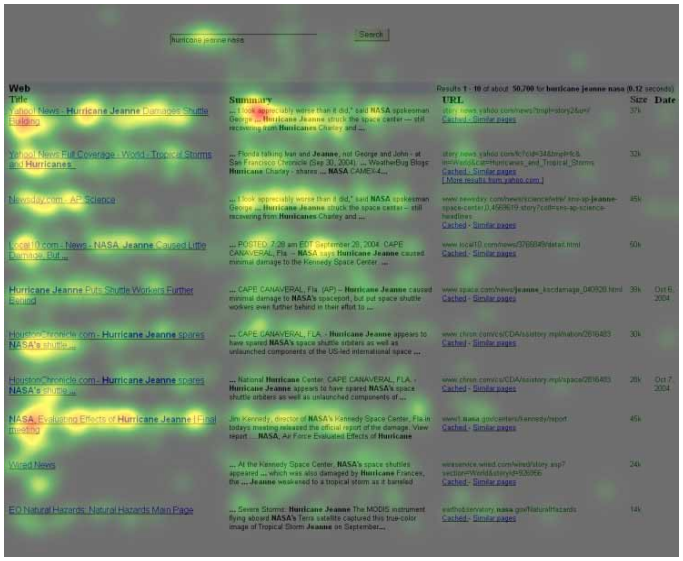

Figure 5. Hotspot plot for an information task on the tabular interface

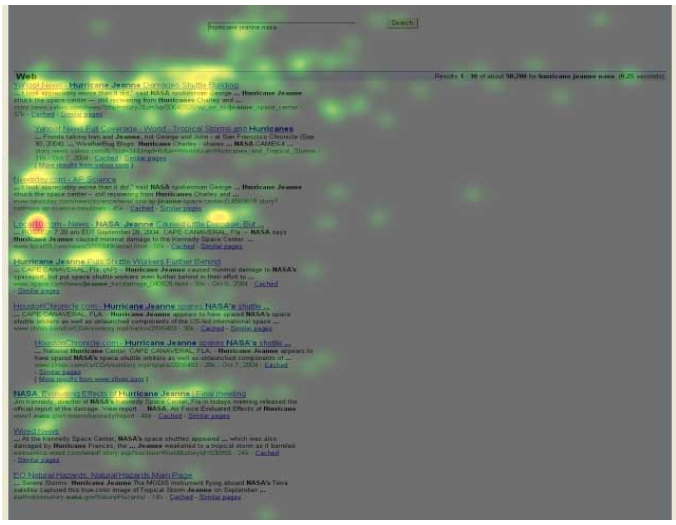

Figure 6. Hotspot plot for an information task on the list interface

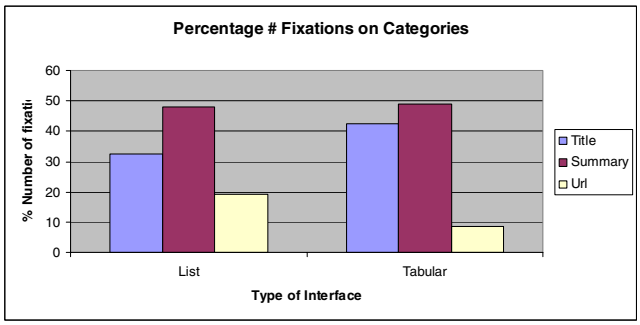

Figure 7. Percentage fixations on different categories for the list and tabular interfaces

The probability of making a transition in the same category of the result was significantly different ($F(1, 60) = 111.32$, $p<0.001$) for list and tabular interface. However, this probability was not significantly different ($F(1, 60) = 0.16$, $p>0.05$) for type of task. Figure 8 shows the mean probability of transition from one category (e.g., Title, Summary, and URL) to another

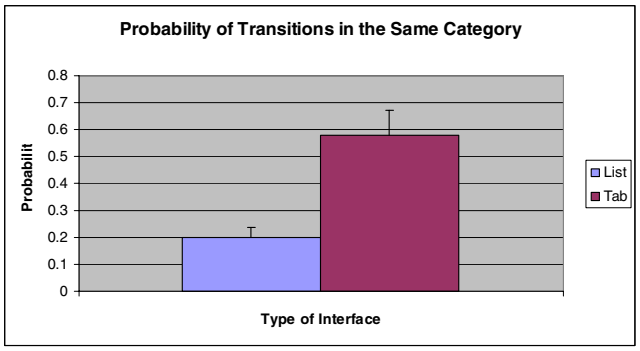

Figure 8. Mean probability of transitions made in the same category for the two interfaces

Subjective measures

The perceived time of task completion ($F(1, 30) = 0.87$, $p>0.05$), perceived accuracy in choosing results ($F(1, 30) = 2.14$, $p>0.05$), preference for the type interface ($F(1, 30) = 2.0$, $p>0.05$), and overall satisfaction ($F(1, 30) = 0.74$, $p>0.05$) did not significantly differ for the two interfaces.

DISCUSSION

Time taken on the two interfaces did not significantly differ, although task time on tabular interface (58.6sec) was marginally longer than on the list interface (42.1sec). This can be attributed to the practice of viewing results on search engines such as Google. The user expertise could not be equalized by just familiarizing users with the tabular interface. Additionally, Duchowski (2002) reports that familiarity of the visual display influences fixation duration, and since fixations contribute to 90% of the viewing time, longer search time on the tabular interface can be attributed to difference in the familiarity with the two displays.

The number of errors or number of click on the wrong results did not significantly differ for the two interfaces, and for the two types of tasks. This indicates that different visual interface designs of search results did not induce an altered clicking behavior in users.

The probability of making a transition to the same category was significantly higher for the tabular interface than for the list interface. This suggests that users preferred to scan a particular category of results on the tabular interface and selectively attend to a particular category due to the vertical arrangement of data, hence showing tendencies to move within columns, rather than between columns. The tabular interface may have allowed users to prioritize elements or categories of the abstract according to their need. No significant difference was found between the probabilities of making same category transitions for the type of task,

suggesting scanning strategy does not change depending on the tasks, while it changes with the type of interface.

The mean fixation durations are content independent measures (Goldberg, 1999) and hence any difference in this metric can be attributed to the interface design. Higher mean fixation durations for list interface (289ms) than those compared to the tabular interface (271ms) is suggestive of a higher cognitive effort on the list interface.

The number of fixations on the summary element of the results was significantly higher for navigation tasks than for the information task. Similar percentages of fixations were found on the summary category for the list (48.12%) and tabular interface (48.92%). The navigation task of finding the homepage of a university that incorporates Stirling engines in its curriculum was found to be difficult by participants, who indicated this concern in the post-task questionnaire. This may have required more careful reading for selection confirmation, hence increasing the number of fixations in the summary category.

Number of fixations in the URL category of results significantly differed for list interface (19.28%) and the tabular interface (8.49 %), suggesting that the list interface led to reading most of everything that was encountered.

CONCLUSION

Users assign weights to different elements of search result's abstract and selectively view information. A top to bottom scanning strategy is evident in both types of interfaces. Overall, users evaluated results based on the Summary that the search engine provided. The eye movements' data supplemented with the conventional usability measures such as time and errors can help in evaluation of search interfaces. Search interfaces can be designed to provide flexibility in the choice of scanning strategies.

FUTURE RESEARCH

Future research is needed to evaluate the list interface with interfaces that allow users to achieve their information search goal non-linearly. Additionally, user-centric interface design can be achieved through studying a wide variety of search tasks scenarios.

ACKNOWLEDEMENTS

The authors would like to thank Sajay Sadasivan for his constructive criticism on this research idea.

REFERENCES

AllTheWeb. (2000). Retrieved on 09/26/2004 from http://www.alltheweb.com

Bandos J., and Resnick M. L. (2004). Improving user search with embedded Boolean search hints. Proceedings of Human Factors Ergonomics Society 48th Annual Meeting, New Orleans, LA, USA, 1523-1527.

Broder A. (2002). A taxonomy of web search. ACM SIGIR Forum, 36 (2): 3-10.

Drori O. (2003). How to display search results in digital libraries-user study. Proceedings of New Developments in Digital Libraries, 13-28.

Duchowski, A. T. (2002). A Breadth-First Survey of Eye Tracking Applications. Behavior Research Methods, Instruments, & Computers (BRMIC), 34 (4): 455-470.

Dumais, S., Cutrell E., and Chen H. (2001). Optimizing Search by Showing Results in Context. Proceedings of ACM CHI Conference, 3 (1): 277-284.

Fallow D. (2005). Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve. Pew / Internet and American Life Project. Retrieved February 2005, from http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf

Forrester report. (October 2000). Retrieved on 09/26/ 2004 from http://www.forrester.com

Goldberg, J. H., and Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. International Journal of Industrial Ergonomics, 24, 631-645.

Granka L., Hembrooke H., Gay G., and Feunser M. (2004). Eye-Tracking Anaysis of User Behavior in WWW Search. Proceedings of the 27th annual international conferenceon Research and development in information retrieval, 478-479.

Klockner K., Wirschum N., and Jameson A. (2004). Depth and Breadth-First Processing of Search Results Lists. Proceedings of ACM CHI Late breaking Results Poster, 1539.

Käki , M. (2005). Findex: Search Result Categories Help Users When Document Ranking Fails. Proceedings of ACM CHI Conference. 131-140.

RealNames. (2000). Retrieved on 09/26/2004 from http://www.realnames.com

Resnick M. L., Maldonado C. A., Santos J. M., and Lergier R. (2001). Modeling On-line Search Behavior Using Alternative Output Structures. Proceedings of the Human Factors and Ergonomics Society 45th Annual Conference, Minneapolis, MN, USA, 1166-1171.

Salogarvi J. Kojo I., Jaana S., and Kaski, S. (2003). Can relevance be inferred from eye movements in information retrieval? Proceedings of the Workshop on Self-organizing Maps, Hibikino, Kitakyushu, Japan, 261-266.

WebTop. (2000). Retrived on 09/26/2005 from http://searchenginewatch.com/sereport/article.php/2163451