Discriminating the Relevance of Web Search Results with Measures of Pupil Size

Flavio T.P. Oliveira[1,2] , Anne Aula[2] , Daniel M. Russell[2]

[1]University of California, Berkeley
Berkeley, CA 94618-1650 USA
oflavio@gmail.com

[2]Google
Mountain View, CA 94043 USA
{anneaula, drussell}@google.com

ABSTRACT

The overwhelming amount of information on the web makes it critical for users to quickly and accurately evaluate the relevance of content. Here we tested whether pupil size can be used to discriminate the perceived relevance of web search results. Our findings revealed that measures of pupil size carry information that can be used to discriminate the relevance of text and image web search results, but the low signal-to-noise ratio poses challenges that need to be overcome when using this technique in naturalistic settings. Despite these challenges, our findings highlight the promise that pupillometry has as a technique that can be used to assess interest and relevance in web interaction in a non-intrusive and objective way.

Author Keywords
Web Search Results, Pupil dilation, Relevance.

ACM Classification Keywords
H5.2. [Information interfaces and presentation]: User interfaces evaluation /methodology

INTRODUCTION

The popularization of the Internet brought many changes to the skill set necessary to be successful in the modern world. The overwhelming amount of information on the web is responsible for one of the most important changes: the acute increase in the need to make quick evaluations of the relevance of content. The task of measuring and understanding the processes underlying user interest and engagement is therefore more important than ever. The elusive challenge lies in accomplishing this task non-intrusively. A widely used approach for measuring the relevance of content is to record subjective self-reports through questionnaires and interviews. This approach is problematic because subjective assessments are cognitively mediated and subject to contextual factors [1]. Researchers have thus looked for more objective and less intrusive measures. To this end, eye-tracking studies have recently become increasingly popular. Studying gaze behavior makes it possible to analyze where users look at on a given User Interface (UI), and to make inferences about the cognitive processes (e.g., perception, comprehension, planning, decision-making and learning) experienced by users while interacting with UIs [2].

Despite the increasing popularity of eye-tracking in Human-Computer Interaction (HCI) research, pupil size—which is recorded by most eye-tracking devices—has received considerably less attention. It is commonly known that pupil size varies to control the amount of light that enters the eye. A less known phenomenon is that pupil size also varies in response to different levels of cognitive engagement and effort [3] and in response to the interest value of stimuli [4]. In addition, recent studies have suggested that pupil size can be used to measure the effects of emotional feedback [5] and as an input modality used to control computer games [6].

Measures of pupil size have the potential to serve as non-intrusive and objective physiological correlates of user interest. This, along with the fact that eye-trackers are increasingly common in usability laboratories, makes pupillometry an intriguing and promising technique for HCI research.

Our question of interest was whether or not pupil size can be used to discriminate the relevance of web search results. We conducted two experiments in which we presented participants with search tasks and then displayed search results that were either relevant or irrelevant to the tasks. We hypothesized that pupil size would differ between the two conditions with the exposure to relevant results eliciting increased pupil size relative to the exposure to irrelevant results.

METHODS AND PROCEDURES

Participants

Twenty-two Google employees (5 women) participated in Experiment 1. Seventeen Google employees (8 women) participated in Experiment 2. The data from 6 participants in Experiment 1 and from 4 participants in Experiment 2 were excluded from the analysis because of technical difficulties or excessive noise in the recordings.

CHI 2009, April 4–9, 2009, Boston, MA, USA.
Copyright 2009 ACM 978-1-60558-246-7/09/04…$5.00.

Procedure

Participants sat in a well-lit room in front of a 17-inch monitor with an integrated eye tracker. We instructed participants that they would be presented with search tasks (e.g., find an image of a clown; find reviews on running shoes) and would then have to rate which of three results presented to them was most relevant to the task. After pressing the space bar to indicate that they had finished reading the search task, participants fixated their vision on a black fixation cross that was presented at the center of the screen on a white background for a variable time period between 4000 and 5000 ms. Participants then saw a single search result (text or image) for 5000 ms. This process repeated for 3 different search results. At the end of this sequence, participants saw a screen containing the previously displayed results. At this moment, participants had to choose which of the three results was most relevant to the search task by pressing the '1','2 'or '3' keys on the keyboard. In 12 of the 24 trials in Experiment 1, the stimuli were Google web search results (with a page title, snippet and a URL) extracted from a search on the topic of the task and in the remaining 12 trials, the stimuli were image thumbnails extracted from Google Image search. In Experiment 2, all the stimuli in the 24 trials were image results. Of the three results presented for each trial, two were irrelevant and one was relevant for the task (as judged by the researchers).

Pupil Diameter Measures

To record pupil diameter, we used a Tobii 1750 eye tracker running Tobii Studio software. The eye tracker was calibrated to each participant's eyes prior to the start of the testing session. Pupil diameter was sampled from both eyes at 50 Hz. We used an automated algorithm to correct for blinks. Blinks were characterized as missing data points. For each blink, we removed the three samples immediately before and after the missing data points. We then linearly interpolated the data. Trials with more than one second of missing data were excluded from further analysis. We averaged the blink-corrected data for the two eyes and then segmented the data into epochs that started 500 ms prior to the onset of each result and lasted until 5000 ms after the onset. We then calculated a baseline value as the average pupil diameter in the 500 ms time window preceding the presentation of each result and subtracted this value from the corresponding epoch to measure the change in pupil diameter elicited by the results. Then, we averaged the epochs separately for the relevant and the irrelevant results.

EXPERIMENT 1: ANALYSIS AND RESULTS

After debriefing the participants and conducting a preliminary analysis of participants' accuracy in the tasks, we decided that one of the text result tasks was ambiguous. Less than half of the participants chose the result that we had considered as the most relevant for this task. We excluded this task from further analysis. The mean accuracy for the remaining text result tasks was 99% (2% standard deviation: SD). The mean accuracy for the image result tasks was 97% (7% SD). To compare the pupillary response elicited by the relevant and irrelevant results, we calculated the difference between those two conditions separately for image and text results by subtracting the values for the relevant results from the values for the irrelevant results. We then calculated 95% bias corrected and accelerated bootstrap confidence intervals (C.I.) for the difference by creating a paired-sample bootstrap distribution of 10,000 resamples for each data point. Next, we calculated the mean difference between the relevant and irrelevant results in the 3000-5000 ms time-window and tested this difference for significance using a one-tailed paired-sample permutation test based on all possible permutations (2^{16}) of the data.

The results showed that the relevant text results elicited increased pupil size during this time-window compared to the irrelevant text results (p = .0045) [**Figure 1**, top plot]. There were no significant differences between relevant and irrelevant image results (p > .5) [**Figure 1**, bottom plot].

Principal Component Analysis (PCA) of Experiment 1

To further investigate the effect of search result relevance on the pupillary response, we conducted principal component analysis (PCA) on the pupil diameter data. The averaged pupil diameter traces [**Figure 1**] represent the sum of multiple factors that influence pupil diameter. Among those factors are the pupillary response to changes in luminance and the pupillary response to stimulus relevance. With the PCA, we were interested in separating those factors into unique components. We first conducted separate PCAs for text results and image results and found similar components in both cases. We therefore chose to conduct a new PCA including both types of results. We computed the PCA on the averaged data per participant and condition. Each data point in time was treated as a dependent variable. This procedure was followed by a Varimax rotation to concentrate each factor on a distinct time period of the pupil diameter trace [7].

A scree plot indicated that three factors derived from the PCA were primarily responsible for the accounted variance. To test for differences between relevant and irrelevant results, we conducted one-tailed paired-sample permutation tests on the factor scores. Factor scores represent a measure of the magnitude of a specific factor in a specific pupil diameter trace. We found significant differences between relevant and irrelevant text results in the factor scores for Factor 1 (means: 0.54 relevant; -0.25 irrelevant, p = .001) and Factor 2 (means: -0.21 relevant; 0.28 irrelevant, p = .018). We found no significant differences in the factor scores for Factor 3 (p > .1). We also found no significant differences in factor scores between relevant and irrelevant image results for all 3 factors (all p > .1).

EXPERIMENT 1: DISCUSSION

Consistent with our prediction, we found that relevant *text results* elicited increased pupil dilation relative to irrelevant text results. This difference emerged at around 3 seconds after the onset of the results and overlasted the duration of the presentation of the results [**Figure 1**, top plot]. This

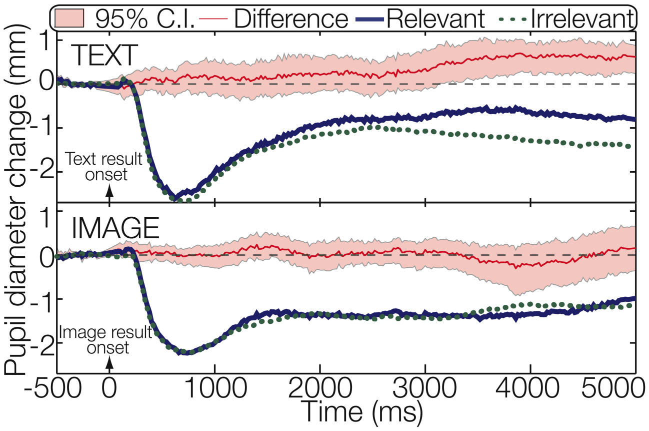

Figure 1. Grand averaged (n = 16) pupil diameter change for Experiment 1. Time zero represents the onset of a text result (top plot) or an image result (bottom plot). The change in pupil diameter is plotted relative to the average of a 500 ms pre-stimulus baseline for relevant (solid blue) and irrelevant (dotted green) results. The pink shaded area displays the 95% confidence interval (C.I.) for the difference (thin red).

finding was further substantiated by the PCA. We found a higher factor score for relevant text results compared to irrelevant text results for Factor 1. This factor seems to represent well the difference between the average pupil diameter traces for relevant and irrelevant text results [**Figure 2**]. We also found differences in Factor 2 scores between relevant and irrelevant text results. This factor peaks at around one second after the onset of a result, suggesting that it may be possible to identify differences in pupil size related to the relevance of text results as early as one second after the onset of the result. We found no differences in the scores for Factor 3, which appears to represent the initial pupillary response to changes in luminance triggered by the onset of the results.

Unlike the pupillary response to text results, we did not find significant differences between relevant and irrelevant *image results* in the averaged pupil diameter traces and in the PCA. Despite the fact that participants' judgments were in high agreement (97%) with our assessment of which images were most relevant to the task, participants' verbal reports during debriefing suggested that the image tasks were more ambiguous. This could have led to a lack of difference in the pupillary response, confounding our results.

EXPERIMENT 2: ANALYSIS AND RESULTS

To address the possibility that the ambiguity of the image tasks confounded the results of Experiment 1, we conducted a second experiment in which the relevance discrimination task was purposely made easier and more objective.

We used similar procedures as those used in Experiment 1 to analyze the data collected in Experiment 2. The mean accuracy on the task was 99% (2% SD).

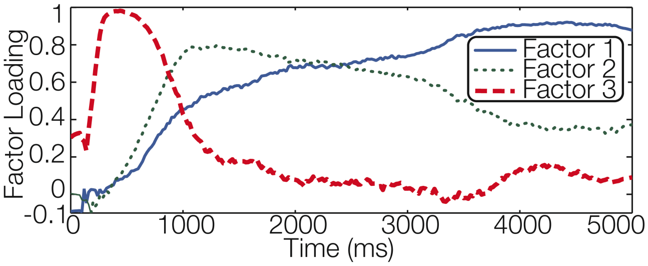

Figure 2. Factor loadings for the three factors extracted from the PCA on the pupil diameter data from Experiment 1. The loadings represent a measure of association between each time point with each factor. The 3 factors accounted for 96.5% of the total variance (Factor 1: 48.0%, Factor 2: 34.6% and Factor 3: 13.8%)

The pupil diameter results showed that unlike in Experiment 1, relevant image results elicited increased pupil size when compared to irrelevant image results (**Figure 3**). We performed a one-tailed permutation test based on all possible permutations (2^{13}) of the data to test the mean difference in pupil diameter between relevant and irrelevant results in the 500-2500 ms time window. The results revealed a highly significant difference (p = .0005).

Principal Component Analysis (PCA) of Experiment 2

The PCA we conducted for Experiment 2 followed similar procedures as those described for Experiment 1. A scree plot indicated that three factors were primarily responsible for the variance in pupil diameter. We found significant differences between relevant and irrelevant results for Factor 2 (means: 0.49 relevant; -0.49 irrelevant, p = .001). We did not find any significant differences for Factor 1 (means: 0.20 relevant; -0.20 irrelevant, p > .1) and Factor 3 (means: 0.05 relevant; -0.05 irrelevant, p > .3) **[Figure 4]**.

EXPERIMENT 2: DISCUSSION

Consistent with our original prediction, we found that relevant image results elicited increased pupil size relative to irrelevant image results in Experiment 2. This difference emerged at around 500 ms and peaked at around 1500 ms after the onset of the results. This early difference seems to be well captured by Factor 2 extracted from the PCA. Unlike in Experiment 1, the late rising of pupil diameter expressed by Factor 1 did not differentiate relevant and irrelevant results. This is possibly due to the fact that the discrimination task in Experiment 2 was considerably simpler than in Experiment 1. This likely led participants to conclude their judgment of relevance sooner after the onset of the result in Experiment 2 than in Experiment 1. This result supports the conclusion that the findings for image results in Experiment 1 may have been confounded by the ambiguity of the tasks and supports the notion that pupil dilation can be used to extract information about the relevance of both text and image results. It also suggests that it may be possible to identify differences in pupil size related to the relevance of web content as early as 500 ms after the onset of the stimulus.

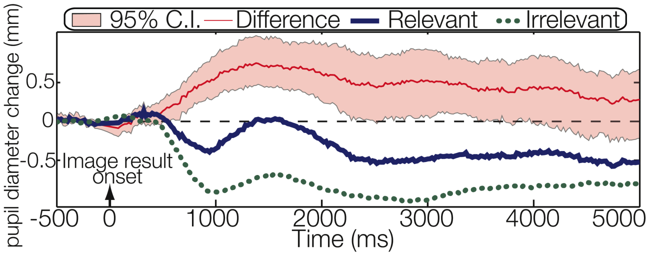

Figure 3. Grand averaged (n = 13) pupil diameter change for Experiment 2.

GENERAL DISCUSSION AND CONCLUSION

Our results demonstrated that measures of pupil size contain information about the relevance of web search results. This was a preliminary investigation in which we tried to control the variables extraneous to the purpose of the experiment as much as possible. Some of the problems we encountered suggest that it may be challenging to use this technique under less controlled conditions.

The first challenge that researchers may face in the applied setting is the sluggishness of the pupillary response. Our analyses on the averaged data suggest that it may take between 500 to 4000 ms to detect differences related to relevance using measures of pupil size. This may not provide enough temporal precision to discriminate which result is most relevant during the rapid evaluation of a search results page. Previous work suggested that users maintain fixation on search results for around 300 to 2000 ms depending on the position of the result [8].

Another challenge that researchers may face is overcoming the low signal-to-noise ratio of pupillometry. This may be an issue when trying to discriminate results that are close together in the continuum of relevance. Here we used search results that lay on opposite extremes of this continuum and it may be more challenging to discriminate results in real search scenarios. Low signal-to-noise may also be an issue when trying to discriminate relevance on a trial-by-trial basis instead of on averaged data such as in the two present experiments.

A third challenge that researchers may face is dealing with differences in luminance between different search results. This may be an even more critical issue when evaluating image results. We dealt with this challenge by averaging together the pupillary responses to several different search results, which made our data less susceptible to individual variations in luminance in the search results. We also conducted PCA on the data, which assisted in separating the pupillary response to changes in luminance from the pupillary response to the relevance of the results.

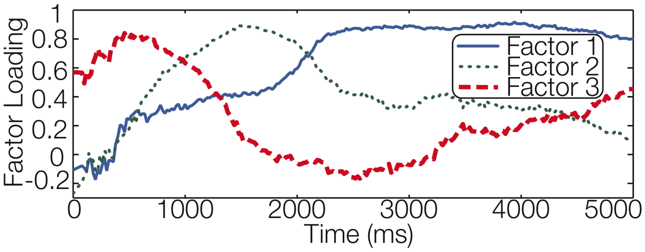

Figure 4. Factor loadings for the three factors extracted from the PCA on the pupil diameter data from Experiment 2. The 3 factors accounted for 92.7% of the total variance (Factor 1: 53.0%, Factor 2: 29.4% and Factor 3: 10.4%)

It is clear that the challenges presented to the applied researcher are not minute. However, the fact that pupil size carries information that can be used to discriminate the relevance of search results, even if under controlled conditions, highlight the promise that pupillometry has as a technique that can be used to assess interest and relevance in web interaction in a non-intrusive and objective way. This should serve as motivation for future work developing novel experimental paradigms and signal processing techniques that address the challenges presented here. This work might prove to be particularly beneficial as eye tracking devices capable of recording pupil size become cheaper and more widely available [9].

REFERENCES

1. Wilson, G.M., Sasse, M.A. Do users know what's good for them? Utilizing physiological responses to assess media quality. In *People and Computers XIV: Proceedings of HCI*, Springer (2000), 327-339
2. Jacob, R. & Karn, K. Commentary on Section 4: Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, Elsevier (2003), 457-470
3. Iqbal, S.T, Zheng, X.S.,Bailey, B.P.. Task-evoked pupillary response to mental workload in human-computer interaction. *Ext. Abstracts CHI 2004,* ACM Press (2004) 1477-1480
4. Hess, E.H., Polt, J.M. Pupil size as related to interest value of stimuli. *Science 132*, 3423 (1960), 349-350
5. Aula, A. & Surakka, V. Auditory emotional feedback facilitates human-computer interaction. *Proc. of HCI* (2002) 337-349.
6. Ekman, I., Poikola, A., Mäkäräinen, M., Takala, T. & Hämäläinen, P. Voluntary pupil size change as control in eyes only interaction. *Proc. of ETRA* (2008), 115-118.
7. Kayser, J., Tenke, C.E. Optimizing PCA methodology for ERP component identification and measurement: theoretical rationale and empirical evaluation *Clinical Neurophysiology 114* (2003), 2307–2325
8. Cutrell, E., Guan, Z. What are you looking for?: an eye-tracking study of information usage in web search. *Proc. CHI 2007*, ACM Press (2007), 407-416
9. Haro, A., Essa, I., Flickner, M. A non-invasive computer vision system for reliable eye tracking *Ext. Abstracts CHI 2000*, ACM Press (2000), 167-168