Differences in eye-tracking measures between visits and revisits to relevant and irrelevant Web pages

Jacek Gwizdka
School of Information
University of Texas at Austin
Austin, TX, USA
sigir2015@gwizdka.com

Yinglong Zhang
School of Information
University of Texas at Austin
Austin, TX, USA
yzhang@utexas.edu

ABSTRACT
This short paper presents initial results from a project, in which we investigated differences in how users view relevant and irrelevant Web pages on their visits and revisits. The users' viewing of Web pages was characterized by eye-tracking measures, with a particular attention paid to changes in pupil size. The data was collected in a lab-based experiment, in which users (N=32) conducted assigned information search tasks on Wikipedia. We performed non-parametric tests of significance as well as classification. Our findings demonstrate differences in eye-tracking measures on visits and revisits to relevant and irrelevant pages and thus indicate a feasibility of predicting perceived Web document relevance from eye-tracking data. In particular, relative changes in pupil size differed significantly in almost all conditions. Our work extends results from previous studies to more realistic search scenarios and to Web page visits and revisits.

General Terms
Experimentation, Human Factors, Measurement.

Keywords
H. Information Systems. H.3 INFORMATION STORAGE AND RETRIEVAL. H.3.3 Information Search and Retrieval. Subjects: Search process.

Author Keywords
Web search; relevance judgment; eye-tracking.

1. INTRODUCTION
Notwithstanding the large body of work on information relevance and relevance judgments, we still do not fully understand the process of how people judge relevance of documents. It is generally agreed that relevance judgment process is dynamic and multi-dimensional [3, 12], but we are only beginning to understand how the various relevance aspects might be applied by users in the process of assessing documents. A formal model of aggregating relevance aspects was proposed [6], but only few research projects attempted to deal with that process from the cognitive perspective [9]. Though, there is also a more recent work in which a cognitively-influenced two-stage model of document assessment was proposed [23].

Given the notion of multi-stage relevance judgment process, we aim to investigate changes in cognitive processes as they are reflected in differences in how documents are viewed and read. In the current project we are looking for evidence of such differences in cognitive processing in Web information search that leads to finding one or more relevant documents. This type of investigation lends itself well to the application of eye-tracking as a data collection method. Eye-tracking helps to get insights into cognitive processes involved in assessing document relevance. Prior work that used eye-tracking in investigation of differences in user behavior between irrelevant and relevant documents tended to use simple documents (e.g., individual sentences or short paragraphs); work that used more realistic tasks and more complex documents (e.g., web pages) did not document visits and revisits, neither did it examine pupil dilation. Prior work tended to look at differences in eye measurements between levels of document relevance, but it did not examine differences at sequential stages in the information search process. Our goal in this short paper is to extend previous work by examining differences in eye-tracking derived measures on Web search tasks conducted on English Wikipedia on pages marked by users as relevant or irrelevant. We examine fixation durations and saccade-related measures, but focus mainly on pupil dilation. To investigate differences between stages of the search process, we operationalize them in a simple way as first visits and subsequent revisits to irrelevant and relevant pages. We formulated the following hypotheses:

H1. Pupil dilation and (some other) eye-tracking measures will differ between relevant and irrelevant pages.

H2. Pupil dilation and (some other) eye-tracking measures will differ between first and subsequent visits to Web pages.

H3. Pupil dilation and (some other) eye-tracking measures will differ between visits to relevant pages when a page relevance was decided compared to other visits to the same relevant pages, when the pages were not judged as relevant yet.

In the next section, we present related work to place our project in context and to motivate our hypotheses.

2. RELATED WORK
One reason why improved understanding of relevance judgment process from user's perspective is useful, is to better model differences between trained relevance judges and actual users. In the already mentioned work [23], Yilmaz et al. proposed a two-stage model of document assessment. The two stages are "1. initial assessment" and "2. extract utility". The authors used this model to explain disparities between judges and users relevance assessments. In addition to being motivated by the notion of stages in document assessment, our work draws on the authors' finding that effort plays an important role in user assessment of document relevance and their argument for including effort in modeling relevance judgment process. Other researchers also

SIGIR '15, August 09 - 13, 2015, Santiago, Chile
 ISBN 978-1-4503-3621-5/15/08…$15.00
DOI: http://dx.doi.org/10.1145/2766462.27677950.

postulated for including effort in modeling interactive information search; for example, using cost of search actions to explain some aspects of search behavior [1], or using search effort to explain search task success [2]. Furthermore, Villa and Halvey [21] showed a relationship between mental effort and relevance levels of judged documents. As we will explain later, pupil dilation, which is typically provided by eye-trackers, is related to mental effort.

Eye-tracking has received considerable attention as a data source useful in information retrieval research. Much of the work has concentrated on eye fixation patterns on ranked search results pages [4,7,14,19]. While this work brought an improved understanding of eye gaze on web pages in general, and on search engine results pages, specifically, it did not address how people view and read documents that differ in their degree of relevance to the user's information need. In other work, eye-tracking techniques have been used in research on relevance judgments. For example, Loboda et al. [16] focused at the word-level relevance and showed a relationship between eye fixation measures and the relevance of terminal words in sentences. Research presented in [20] and [5] demonstrated that user's eye movement behaviors would shift between scanning, reading, or deciding processes, and that reading and skimming behavior was heavily influenced by relevance (especially by topicality of text). In agreement with these findings, [8,10] reported that relevant and irrelevant paragraphs and short documents, respectively, could be discriminated based on gaze data and eye-movement patterns in reading experiments. Although eye-tracking derived measures were investigated in previous research, most of the above mentioned studies focused on fixation duration, fixation counts, and fixation locations; pupil dilation has been generally neglected. Cognitive psychology research informs us that under constant illumination pupil dilation is associated with mental effort and attention [18] and thus, plausibly, it may be related to cognitive processing of documents. Indeed, Gwizdka and Oliveira et al. [10,17] demonstrated that changes in pupil size are related to levels of relevance – pupil dilated for more relevant stimuli, both for text and images. In light of this evidence, it seems reasonable to incorporate pupilometry in interactive information retrieval research. The use of pupilometry is additionally corroborated by the relationship between pupil dilation and mental effort.

In summary, eye-tracking is a feasible approach to examining user's implicit relevance judgments. With the sizable body of prior work, there is still a need for more research conducted on more realistic search task scenarios and documents (e.g., on the live Web), and paying attention to sequential stages of search and to a wider range of eye-tracking measures. That's where we position our current work.

3. METHOD

We conducted a lab-based experiment of Web search on Wikipedia. 32 native English speakers (15 females) with a normal, to corrected-to-normal vision, attended individual experiment sessions held in the Information eXperience (IX) lab in the School of Information at University of Texas at Austin. Each session was completed within 1.5 hours; participants received $30 for their participation. Each participant was asked to complete four search tasks that were designed to differ in complexity (within-subject design). The searches were conducted on Wikipedia using a commercial test search engine created by Search Technologies Corp. We used the commercial search engine, because Wikipedia does not provide full-text search. Participants used a PC computer running Windows 7 that was equipped with Tobii T60 eye-tracker. The eye-tracking and interaction data was collected using Attention Tool software [13], while the task rotations and questionnaires were controlled by YASFIIRE software [22]. Task rotations were assigned to participants in a random order. In each task, participants read task description, completed pre- and post- task questionnaires, and searched Wikipedia. At the end of the session, they answered an exit questionnaire. There were no time limits set for search tasks. In this paper, our focus is on viewing behavior of Wikipedia pages. During a visit to a Wikipedia page, in addition to reading the page, a user could open a search task description on screen, bookmark the page and enter notes, edit the previously entered notes, and decide to complete the task by entering and editing the final notes about the results of their search. According to the instructions given to our participants, bookmarking a page signified that the page was perceived by a participant as relevant to the search task, thus the bookmarking events provided us with binary relevance judgments. Users were also able to delete any bookmark, if they decided that a page was not relevant any more. The visits to Wikipedia pages (relevant and irrelevant) were categorized into first visits and revisits. We further distinguished these visits to relevant pages when relevance judgments were made (these included first visits and revisits). We isolated time periods when a user was viewing a Wikipedia page (that is not reading task description nor entering notes) and considered each of them separately in our data analysis. We call user's activity during these time periods a *page viewing state* and emphasize that a visit to a Web page may contain one or more such states. Extraction of isolated page viewing states serves a purpose of separating cognitive activities related to document processing from other activities, and, thus, could be considered as contributing to eye-tracking data cleaning.

The perceived binary level of page relevance, page first visit, revisit, and visit with relevance judgment were factors used in data analysis. Table 1 shows eye-tracking derived variables.

Table 1. Eye-tracking measures

Variable	Description
Fixation duration	Duration of an eye fixation, in milliseconds
Saccade duration	Duration of a saccade, that is of a fast eye movement between eye fixations, in milliseconds
Saccade length	Length of a saccade, in pixels
Saccade angle	Angle of a saccade relative to the horizontal axis, in degrees
Relative pupil dilation	The relative change in pupil diameter: A difference between pupil size at a time *t* and the average pupil size for a participant, normalized by that average

4. DATA ANALYSIS AND RESULTS

We first removed page viewing states with few eye fixation data points by deleting states shorter than 6 seconds. This resulted in removing less than 6% of data. All further analyses were conducted after these short states were removed.

Table 2. Counts of page viewing states

Level	Page and visit types		Count	
1	Irrelevant page	first visit	306	215
2		revisit		91
n/a	Relevant page	first visit	765	358
n/a		revisit		407
3	Relevant page*	first visit*	697	323
4		revisit*		374
5	Rel. page visit with relevance judgment		68	

* shows counts after removing visits when relevance judgments were made, which are contained in level 5

Our data did not satisfy analysis of variance assumptions, therefore we performed non-parametric tests. We constructed a 5-level factor variable (shown in Table 2 & 3) from visits to irrelevant and relevant pages and different visit types (first visits, revisits, and visits to relevant pages with relevance judgment). The Kruskal-Wallis rank sum test conducted with this factor showed significant differences for fixation duration ($\chi(4)^2$=24.05, p<.001), pupil dilation ($\chi(4)^2$=891.33, p<<.0001), saccade duration ($\chi(4)^2$=171.3, p<<.0001), and saccade length ($\chi(4)^2$=41.15, p<.0001).

Table 3. Pupil dilation (mean(SD))

Level	Page and visit types		Pupil dilation	
1	Irrelevant page	first visit	-0.045 (0.054)	-0.0497 (0.053)
2		revisit		-0.0362 (0.054)
3	Relevant page*	first visit*	-0.036 (0.059)	-0.0428 (0.055)
4		revisit*		- 0.0319 (0.063)
5	Rel. page visit with relevance judgment		- 0.033 (0.059)	

* shows counts after removing visits when relevance judgments were made, which are contained in level 5

The post-hoc pairwise comparisons (all checked at p<.05) indicated significant differences in relative change in pupil size in almost all cases, except between revisits to irrelevant and relevant pages and visits when relevance judgment took place; pupil dilation was the largest on these visits, while the smallest dilation was on first visits to irrelevant pages. For other variables, the pairwise comparisons indicated significant differences in only in a few cases. A significantly longer fixation durations on revisits to relevant pages compared with revisits to irrelevant and first visits to relevant pages. Saccade duration on the first visits to relevant pages tended to be longer than on irrelevant pages. The slowest saccades were on relevance judgment visits to relevant pages. The pixel-length of saccades was significantly longer on all visits to irrelevant pages compared with first visits to relevant pages. The length of saccades was significantly shortest on first visits to relevant vs. revisits or relevance judgment visits to these pages. These results generally confirm our hypotheses, though, to a different extent for different variables; their pattern of significance is shown in Table 4.

Table 4. Summary of pairwise comparisons

	Comparisons between 5-level factor levels									
Variable	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**
Fixation duration						+		+		
Saccade duration		+		+			+	+	+	+
Saccade length		+			+			+	+	
Saccade angle										
Relative pupil dilation	+	+	+	+	+	+		+	+	

+ indicates significant pairwise comparison.

For numbering of the 5-level factor levels please see Table 2.

One can expect significant cognitive processes to take place shortly before a decision is made. Therefore, we separately examined the last 2 seconds period of each page viewing state and compared differences between all visits with relevance judgment visits. The Wilcoxon rank sum test showed a significant difference for pupil dilation (W = 479178, p<<.0001). The same test also showed a significant difference between the first and the last two seconds of a visit to a relevant page during with relevance judgment (W = 45615.5, p<0.0001). Other variables were not significantly different. In both cases, pupil tended to be larger during the two second period before the relevance decision was made.

To further investigate how eye-tracking data reflects relevance judgments, we employed classification. In the analysis, we proposed two models to classify two different situations. The first model was used to classify first visits to relevant pages and visits to relevant pages during which participants judged relevance. The second one was used to classify visits to irrelevant pages and visits to relevant pages during which participants judged relevance. Because of the class imbalance in our datasets, we used up-sampling [15] to simulate additional data points, thereby improving balance across classes. For both models, we used flexible discriminant analysis (FDA) and all five variables were included as predictors (Table 1). The parameters of the two models were tuned to maximize Receiver Operating Characteristic (ROC) and cross-fold validation was used to evaluate the performance of the models. Outcomes of both our classification models showed that the relative pupil dilation was the most important predictor (Table 5). The models' performances are shown in Table 6.

Table 5. Variable importance for two classification models

Variable	Model 1	Model 2
Fixation duration	29.38	29.66
Saccade duration	81.27	59.94
Saccade length	46.70	0.00
Saccade angle	0.00	54.94
Relative pupil dilation	100.00	100.00

Table 6. Classification model performance

Model	Accuracy	Sensitivity	Specificity
Model 1	0.57	0.57	0.57
Model 2	0.61	0.57	0.62

5. DISCUSSION AND CONCLUSIONS

The findings show that several eye-tracking derived measures significantly differ between user visits and revisits to relevant and irrelevant Wikipedia pages.

In particular, a larger pupil dilation on visits to relevant pages indicates, in part, a higher mental effort and attention paid to relevant pages, and particularly so on revisits or when a relevance judgment was finally made. This result is inline with the results presented in [10, 21], but it also extends them, since our investigation distinguished first and subsequent visits to pages.

Average pupil dilation during page viewing states did not differentiate between revisits and relevance judgment visits to relevant pages. However, we showed that the significantly larger pupil dilation in the last 2 seconds before the relevance judgment was made, should enable one to differentiate the relevance judgment visits from any other visits to Web pages. This result can be plausibly explained by increased attention during relevance judgment.

Results with respect to fixation duration, saccade duration and length, generally confirm prior work. For example [5, 10], where a tendency for visual scanning of irrelevant pages and continuous reading of relevant pages was shown. Our results indicate that on revisits to relevant pages and when relevance judgments were made, the Web pages were read more carefully.

Previous studies that employed eye-tracking in characterization of text relevance were typically conducted under more constrained human-information interaction and, for example, examined only reading of prepared sentences [20], paragraphs [8] or short text documents [10] and not user-selected Web pages. Our work

extends the prior research by investigating more realistic and complex interactive information retrieval scenarios on the Web and by conducting analysis at the level of page viewing states, which we constructed to isolate cognitive activities not related to document viewing/reading.

The classification results demonstrate a promise in using eye-tracking variables in predicting user's perceived relevance of Wikipedia pages. While accuracy of the two models shown in this paper is rather modest, it could be improved by combining with other interaction data and by applying further processing to pupil size, following, for example, approaches described in [11, 17].

One of the motivations of our work was the notion multi-stage of relevance judgment, the differences in variables between the visits and revisits provide some indirect support for it. Interestingly, patterns of pairwise comparison significance (Table 4) differ in all but two cases (1-3 and 3-5). Plausibly one could use different combinations of variables to infer different types of page viewing states (Table 2). This finding is a likely indication of differences in cognitive processes, such as shifts in mode of reading (e.g., scanning, skimming, and continuous reading) and levels of mental workload and attention changing between visits and revisits to pages at different levels of relevance.

A limitation of the current analysis is that we did not use time sequence analysis that would allow for more direct modeling of a sequence of stages in document assessment. In the future, we plan to apply techniques such as Hidden-Markov Modeling (HMM) to our data.

Other limitations include, a bias in our data towards relevance as well as use of only one web site, the English Wikipedia. We believe that our results should generalize to other text-heavy web pages and in our future work we plan to broaden the set of web pages we use and to address other limitations. In addition, our analysis was conducted at an aggregate level across all users and all tasks. In the follow up work, we will also examine data at an individual user and task levels.

6. ACKNOWLEDGMENTS

This research was supported, in part, by IMLS Career Development Grant #RE-04-11-0062-11A to Jacek Gwizdka.

7. REFERENCES

[1] Azzopardi, L. 2014. Modelling Interaction with Economic Models of Search. *Proceedings of SIGIR'2014* (New York, NY), 3–12.

[2] Bailey, E. and Kelly, D. 2011. Is amount of effort a better predictor of search success than use of specific search tactics? *Proceedings of the American Society for Information Science and Technology*. 48, 1, 1–10.

[3] Borlund, P. 2003. The concept of relevance in IR. *JASIST*. 54, 10, 913–925.

[4] Brumby, D.P. and Howes, A. 2008. Strategies for Guiding Interactive Search: An Empirical Investigation Into the Consequences of Label Relevance for Assessment and Selection. *Human–Computer Interaction*. 23, 1, 1–46.

[5] Buscher, G. et al. 2012. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Trans. Interact. Intell. Syst.* 1, 2, 9:1–9:30.

[6] Da Costa Pereira, C. et al. 2012. Multidimensional relevance: Prioritized aggregation in a personalized Information Retrieval setting. *Information Processing & Management*. 48, 2, 340–357.

[7] Cutrell, E. and Guan, Z. 2007. What are you looking for?: an eye-tracking study of information usage in web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY), 407–416.

[8] Fahey, D. et al. 2011. Document Classification on Relevance: A Study on Eye Gaze Patterns for Reading. *Neural Information Processing*. B.-L. Lu et al., eds. Springer Berlin Heidelberg. 143–150.

[9] Greisdorf, H. 2003. Relevance thresholds: a multi-stage predictive model of how users evaluate information. *Information Processing & Management*. 39, 3, 403–423.

[10] Gwizdka, J. 2014. Characterizing Relevance with Eye-tracking Measures. *Proceedings of the 5th Information Interaction in Context Symposium* (New York, NY), 58–67.

[11] Hossain, G. and Yeasin, M. 2014. Understanding Effects of Cognitive Load from Pupillary Responses Using Hilbert Analytic Phase. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Jun. 2014), 381–386.

[12] Huang, X. and Soergel, D. 2013. Relevance: An improved framework for explicating the notion. *JASIST*. 64, 1, 18–35.

[13] iMotions 2014. *Attention Tool Biometric Research Platform: Version*. iMotions, Inc.

[14] Joachims, T. et al. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Trans. Inf. Syst.* 25, 2.

[15] Kuhn, M. and Johnson, K. 2013. *Applied Predictive Modeling*. Springer Science & Business Media.

[16] Loboda, T.D. et al. 2011. Inferring word relevance from eye-movements of readers. *Proceedings of the 16th international conference on Intelligent user interfaces* (New York, NY), 175–184.

[17] Oliveira, F.T.P. et al. 2009. Discriminating the relevance of web search results with measures of pupil size. *Proceedings of the 27th ACM international conference on Human factors in computing systems CHI'2009* (Boston, MA), 2209–2212.

[18] Onorati, F. et al. 2013. Characterization of affective states by pupillary dynamics and autonomic correlates. *Frontiers in Neuroengineering*. 6, 9.

[19] Pan, B. et al. 2004. The determinants of web page viewing behavior: an eye-tracking study. *Proceedings of the 2004 symposium on Eye tracking research & applications* (New York, NY), 147–154.

[20] Simola, J. et al. 2008. Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*. 9, 4, 237–251.

[21] Villa, R. and Halvey, M. 2013. Is relevance hard work?: evaluating the effort of making relevant assessments. *Proceedings of SIGIR'2013* (New York, NY, US), 765–768.

[22] Wei, X. et al. 2014. YASFIIRE: Yet Another System for IIR Evaluation. *Proceedings of the 5th Information Interaction in Context Symposium* (New York, NY), 316–319.

[23] Yilmaz, E. et al. 2014. Relevance and Effort: An Analysis of Document Utility. *Proceedings of CIKM'2014* (New York, NY), 91–100.