Towards Inferring Web Page Relevance – An Eye-Tracking Study

Jacek Gwizdka, University of Texas at Austin
Yinglong Zhang, University of Texas at Austin

Abstract
We present initial results from a project, in which we examined feasibility of inferring web page relevance from eye-tracking data. We conduced a controlled, lab-based Web search experiment, in which participants conducted assigned information search tasks on Wikipedia. We performed analyses of variance as well as employed classification algorithms in order to predict user perceived Web page relevance. Our findings demonstrate that it is feasible to infer document relevance from eye-tracking data on Web pages. The results indicate that eye fixation duration, pupil size and the probability of continuing reading are good predictors of Web page relevance. Our work extends results from previous studies of text document search that were conducted under more constrained human-information interaction.

Keywords: Information Relevance; Eye-tracking; User Study
Citation: Gwizdka, J., Zhang, Y. (2015). Towards Inferring Web Page Relevance – An Eye-Tracking Study. In *iConference 2015 Proceedings*.

Acknowledgements: This research project was funded, in part, by the IMLS Career Development Grant #RE-04-11-0062-11A.
Contact: iconf@gwizdka.com, ylzhang@utexas.edu

1 Introduction

Understanding a person's information need in search is the wholly grail of interactive information retrieval. Relationship between information need and documents, has been conceptualized as information relevance (e.g., Saracevic, 2007). The last two decades brought a much better understanding of this concept, yet the discussion on the nature of relevance and its underlying factors continue (Huang & Soergel, 2013). In spite of understanding an importance subjective factors play in relevance judgments (e.g. Borlund, 2003; Cosijn & Ingwersen, 2000), relatively little is known how people cognitively process relevant vs. irrelevant texts, and how documents at different degrees of relevance are read. Our research aims to bridge this gap, and, on the applied side, to contribute to the design of better search engines.

2 Background and Related Work

Eye tracking has received considerable attention as a data source useful in information search and retrieval research. Much of the work using eye tracking data has concentrated on eye fixation patterns on ranked search results pages (Brumby & Howes, 2008; Granka, Joachims, & Gay, 2004; Joachims et al., 2007; Pan et al., 2004). For example, Guan & Cutrell (Guan & Cutrell, 2007) examined task type influences on user search behavior by manipulating the positions of target results in navigational and informational tasks. 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 read documents that differ in their degree of relevance to the user's information need.

Under constant illumination pupil dilation has been associated with mental effort and attention (Onorati, Barbieri, Mauri, Russo, & Mainardi, 2013). The work that examined pupil size in relation to information relevance includes (Oliveira, Aula, & Russell, 2009), who found that for text documents and images pupil dilated for more relevant stimuli. In recent work (Gwizdka, 2014) found a similar effect for text documents. He also examined reading patterns on documents of different degrees of relevance, but did so only on constrained human-text interaction.

Work presented in this short paper aims to bridge the gaps identified in past work and to extend previous research.

3 Method

We conducted a controlled, lab-based experiment of Web search on Wikipedia. Participants (N=32; age: from 18 to 37; 15 females; 17 males), who were native English speakers and had a normal, to corrected-to-normal vision, attended individual experiment sessions held in the Information eXperience lab in the School of Information at University of Texas at Austin. Each session was completed within 1.5 hours or less; participants received $30 for their participation. Participants were asked to complete four search tasks that were designed to be at two complexity levels: simple and complex. The searches were conducted using a commercial test search engine created by Search Technologies Corp. with two variations of user interface (UI) created by us. In this paper we focus on user interaction with Wikipedia pages, and therefore we will not be discussing the search engine interfaces. The experiment had a within-subject design with each participant conducting four search tasks at two level of complexity. 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. Participants were asked to save the pages they considered as relevant to the search task and to add notes to these pages. At the end of a session, they answered exit questionnaire. There were no time limits set for search tasks.

We used Tobii T60 remote eye-tracker (Tobii Technology, Stockholm, Sweden). The eye-tracking and interaction data was collected using Attention Tool software (iMotions, 2014), while the task rotations and questionnaires were controlled by our own YASFIIRE software (Wei, Zhang, & Gwizdka, 2014) that controlled task rotations and the questionnaire data collection.

3.1 Independent and Dependent Variables

The independent variable was participant's perceived binary relevance of Wikipedia pages. This measure was obtained by considering Wikipedia pages that were saved and annotated by participants as relevant, and Wikipedia pages that were visited but not saved by participants as irrelevant. Dependent variables were obtained directly or derived from data captured by the eye-tracker. They included eye-fixation durations, and pupil diameter. To accommodate individual differences in pupil sizes, we normalized pupil diameter by calculating its relative changes for each participant; we followed the procedure described in (Xu, Wang, Chen, & Choi, 2011).

We also used eye-tracking measures aggregated per Wikipedia page visit and calculated average fixation duration (AvgFixDur), average changes in pupil size (AvgNormPupil), the number of eye-fixations on a page normalized by duration of the page visit (FixCountPerTime). We obtained additional measures by categorizing eye-fixation data into two reading states: reading in line and scanning; the procedure is described in (Cole et al. 2011). We calculated probability of remaining in the scanning state (pSS) and probability of remaining in the reading state (pRR) for each Wikipedia page.

3.2 Research Questions and Hypotheses

Informed by prior work, cited in section 2, we expected eye-tracking based measures to be related to participants' perceived web page relevance. Thus, we formulated the following hypotheses:

H1: Fixation durations and changes in pupil size on Wikipedia pages differ between relevant and irrelevant pages.

H2: Fixation durations and changes in pupil size shortly before (2 seconds) participants made relevance judgments of Wikipedia pages differ between relevant and irrelevant pages.

H3: Average fixation duration, average changes in pupil size, number of fixations per time, probability of remaining in a scanning state, and probability of remaining in a reading state while visiting Wikipedia pages differ between relevant and irrelevant pages.

H4: Eye-tracking data (average fixation duration, average changes in pupil size, number of fixations per time, probability of remaining in a scanning state, and probability of remaining in a reading state) can be used to predict participants' relevance judgments.

4 Data Analysis and Results

Our data analysis included two phases. In the first phase, we used a series of one-way ANOVAs to test the differences between relevant and irrelevant pages on participants' eye movements. We found significant differences between the relevant and the irrelevant pages (confirming *H1* and *H2*) as shown in Table 1.

Eye Movement Variables	Relevant Page	Irrelevant Page	F-test (ANOVA)

	Mean (sd)	Mean (sd)	
Fixation Duration – whole [ms]	203.10 (117.10)	199.70 (106.70)	15.70*
Changes in pupil size – whole	-0.028 (0.063)	-0.040 (0.055)	701.80*
Fixation Duration – 2 seconds before relevance judgment [ms]	236.90 (166.40)	223.00 (161.30)	4.84*
Changes in pupil sizes – 2 seconds before relevance judgment	-0.013 (0.060)	-0.020 (0.067)	10.61*

Table 1. Descriptive statistics and ANOVA F-test. (Variables labeled as “whole” are measured for the whole duration of visits to a Wikipedia pages, while variables labeled as “2 seconds” are measured for the last 2 seconds before a participant’s relevance judgment. * denotes significance at the p<.05 level.)

Eye Movement Variables	Relevant Page Mean (sd)	Irrelevant Page Mean (sd)	F-test (ANOVA)
Average fixation duration (AvgFixDur) [ms]	205.40 (40.11)	203.60 (40.350)	0.319
Average changes in pupil size (AvgNormPupil)	-0.024 (0.033)	-0.038 (0.036)	23.120*
Number of fixations per time (FixCountPerTIme)	0.002 (0.001)	0.0026 (0.001)	71.640*
probability of continuing to scan (pSS)	0.457 (0.215)	0.448 (0.254)	0.228
probability of continuing to read in line (pRR)	0.970 (0.044)	0.955 (0.056)	11.700*

Table 2. Descriptive statistics and ANOVA F-test for variables aggregated per page. * denotes significance at the p<.05 level)

Table 2 shows the results of ANOVA conducted on eye-tracking measures aggregated per Wikipedia page visit. The results indicated that the relevant and the irrelevant pages had significant differences on the average changes in normalized pupil size, the number of fixations per time, and the probability of remaining in a reading state (confirming *H3*).

The ANOVA analyses demonstrated that variables derived from eye-tracking data differed significantly between the relevant and the irrelevant pages. In the second phase, we attempted to explore the feasibility of using eye-tracking data to predict relevance by employing machine learning approach.

In this part, we applied a classification algorithm to eye-tracking measures collected on 700 independent visits to Wikipedia pages. Five predictors were included: number of fixations per time, average changes in pupil size, average fixation duration, probability of remaining in a scanning state, and probability of remaining in a reading state. Each page was labeled as belonging to one of the two classes: “relevant” or “irrelevant” (65% relevant and 35% irrelevant). Using a stratified random sampling to split dataset, we selected 80% of the pages as training dataset, and the rest were held out for testing (evaluation). We used flexible discriminant analysis (FDA) to predict relevance of documents and parameters of the model were tuned to maximize Receiver Operating Characteristic (ROC). In model training, cross-fold validation (10 fold) was employed to evaluate the performance of FDA and the parameters of the model were determined in terms of the area under the ROC[1].

In model evaluation, we used holdout set to test the FDA model. For this model, the area under the ROC curve was 0.663 (ranging from 0.572 to 0.754), as shown in Figure 1. Its overall accuracy was 62% (ranging from 53% to 70%). The recall (Sensitivity) and the precision (Positive Predictive Value) were 0.756 and 0.687, respectively. The F-measure (the harmonic mean of precision and recall) was 0.720; this result indicated a large overlapping between the true and the estimated classes for our model. The specificity of this model was 0.367. The importance of each predictor in the final model is shown in Figure 1. Based on this result, the number of fixations per time, the average changes in the pupil size, and the probability of remaining in a reading state all played important roles in the prediction of page relevance.

[1] For a perfect model, area under the ROC curve would be 1.

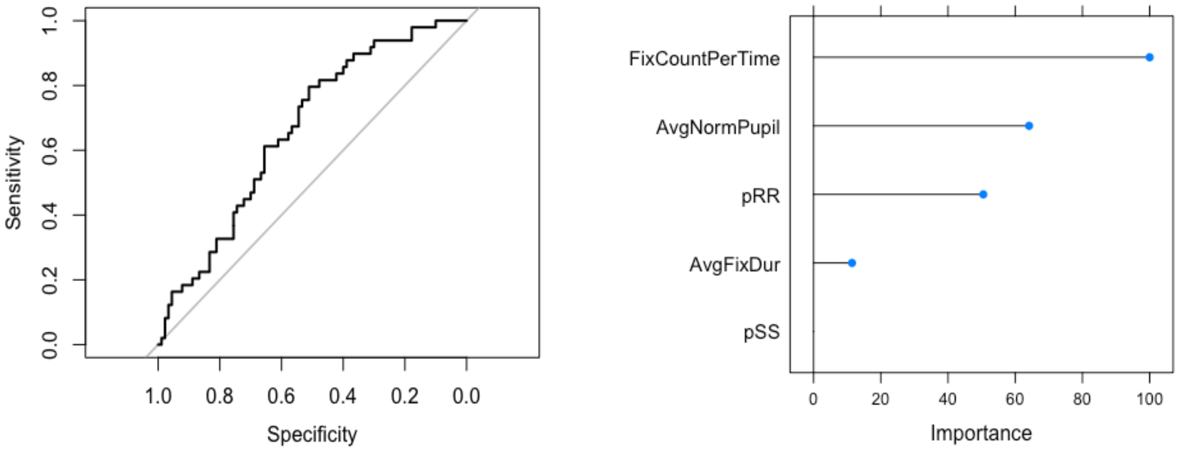

Figure 1. A ROC curve for FDA model and the relative importance of predictors in classification

5 Discussion and Conclusion

Our results show that several eye-tracking derived measures significantly differ between user visits to relevant vis-à-vis irrelevant Wikipedia pages. The classification results indicate that eye fixation duration, pupil size and the probability of continuing reading are good predictors of Web page relevance. Thus, the results demonstrate a feasibility of predicting user's perceived relevance of Wikipedia pages, which confirms our last hypothesis (*H4*). Hence, all hypotheses that we formulated were confirmed.

While our results generally agree with previous work, our main contribution lies in obtaining these findings in an interactive Web information search scenario. 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 text documents and not user-selected web pages (e.g., Gwizdka, 2014).

Limitations of our work include, a bias in our data towards relevance as well as use of only one web site, the English language 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.

References

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925. doi:10.1002/asi.10286

Brumby, D. P., & 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. doi:10.1080/07370020701851078

Cole, M. J., Gwizdka, J., Liu, C., Bierig, R, Belkin, N. J., Zhang, X. (2011). Task and User Effects on Reading Patterns in Information Search. Interacting with Computers, 23(4), 346 - 362. DOI:10.1016/j.intcom.2011.04.007

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, *36*(4), 533–550. doi:10.1016/S0306-4573(99)00072-2

Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 478–479). New York, NY, USA: ACM. doi:10.1145/1008992.1009079

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 417–420). New York, NY, USA: ACM. doi:10.1145/1240624.1240691

Gwizdka, J. (2014). Characterizing Relevance with Eye-tracking Measures. In *Proceedings of the 5th Information Interaction in Context Symposium* (pp. 58–67). New York, NY, USA: ACM. doi:10.1145/2637002.2637011

Huang, X., & Soergel, D. (2013). Relevance: An improved framework for explicating the notion. *Journal of the American Society for Information Science and Technology*, *64*(1), 18–35. doi:10.1002/asi.22811

iMotions. (2014). Attention Tool Biometric Research Platform: Version (Version 5.2). Cambridge, MA.: iMotions, Inc.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Trans. Inf. Syst.*, *25*(2). doi:10.1145/1229179.1229181

Oliveira, F. T. P., Aula, A., & Russell, D. M. (2009). Discriminating the relevance of web search results with measures of pupil size. In *Proceedings of the 27th international conference on Human factors in computing systems* (pp. 2209–2212). Boston, MA, USA: ACM. doi:10.1145/1518701.1519038

Onorati, F., Barbieri, R., Mauri, M., Russo, V., & Mainardi, L. (2013). Characterization of affective states by pupillary dynamics and autonomic correlates. *Frontiers in Neuroengineering*, *6*, 9. doi:10.3389/fneng.2013.00009

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K., & Newman, J. K. (2004). The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 symposium on Eye tracking research & applications* (pp. 147–154). New York, NY, USA: ACM. doi:10.1145/968363.968391

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144. doi:10.1002/asi.20681

Wei, X., Zhang, Y., & Gwizdka, J. (2014). YASFIIRE: Yet Another System for IIR Evaluation. In *Proceedings of the 5th Information Interaction in Context Symposium* (pp. 316–319). New York, NY, USA: ACM. doi:10.1145/2637002.2637051

Xu, J., Wang, Y., Chen, F., & Choi, E. (2011). Pupillary Response Based Cognitive Workload Measurement under Luminance Changes. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2011* (Vol. 6947, pp. 178–185). Springer Berlin / Heidelberg. Retrieved from http://www.springerlink.com.proxy.libraries.rutgers.edu/content/5255211747545701/abstract/

Table of Figures

Table of Tables