Tracking Information Relevance

Jacek Gwizdka, School of Information, University of Texas at Austin; Michael Cole, Rutgers University

Corresponding author: Jacek Gwizdka: neurois2014@gwizdka.com

Relevance is a fundamental concept in information retrieval. We consider relevance from user perspective (Borlund, 2003) and focus on topical relevance that is described as the extent to which the content of a document (or a set of documents) matches the topic of the query or a user's information need (Saracevic, 2007). Judgments of document relevance are important events during user interaction with a search system. Direct and non-intrusive detection of relevance judgements would provide an objective means to capture this important aspect of user's interaction with information while a user is engaged in search. We continue our work that aims to better understand the concept of relevance in order to better support searchers on their tasks. We report on a project in which we explored the possibility of inferring relevance from eye-tracking data and in which as we asked if reading documents that vary in a degree of relevance to a user's information search goal imposes different cognitive demands.

We conducted an experiment, in which participants were asked to find information in short text documents containing news stories. The experiment was conducted in a usability lab equipped with Tobii T-60 eye-tracker. The experimental design shown in Figure 1 is essentially the same as in our fMRI study reported at NeuroIS'2013 (Gwizdka, 2013). We describe it briefly below. Each participant performed two types of tasks: 1) target word search (WS task), and 2) a simulated information search (IS task). Each session included 21 pseudo-randomized trials of each task type, as well as a few training trials. We focus on reporting results from IS task. The IS task involved finding relevant factual information in news stories. First, general task instructions were presented on screen for 30 seconds, next a fixation screen appeared for 4 seconds, then a question was displayed for 8 seconds. The question instructed participants what information they were expected to find in subsequently presented documents. Twenty one questions were presented in pseudo-randomized order, each followed by three news stories of varied relevance: irrelevant – I, partially relevant texts that were on a question's topic, but did not contain the answer – T, and relevant texts that contained the answer – R (Figure 1). Fixation screens were presented for 4 seconds before each text. In addition, to remind participants of the current question, it was repeated briefly (4s) before the second and third text (shown in Figure 1 as "+" above the stimuli). Participants responded by explicitly judging document relevance of sixty-three news stories on a binary scale (yes/no). Before the actual task began, participants performed a few training trials.

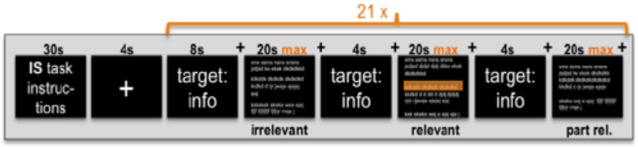

Figure 1. Experimental Design

The documents used in the study came from a large set of news stories obtained from the AQUAINT corpus (Graff, 2002). All news stories were in English and originated from several international sources, such as Associated Press, New York Times, and Xinhua. We selected a subset of stories aiming to achieve a relatively low variation in the text length. We obtained relevance assessments for the documents from TREC Q&A task from 2005 (Voorhees & Dang 2005). We further manually verified the relevance assessments for the selected document subset. The average length of the documents was 178 words (SD=30).

Eye movements were analysed using our reading modelling approach described in (Cole et al., 2011a; 2011b), which is briefly summarized below. Our approach is influenced by the E-Z Reader model (Reichle et al, 2006; Rayner et al, 2011). The assumptions of that model are as follows: 1) reading is serial and words are processed one at a time in the order of their appearance in text, 2) more than one word can be processed on a fixation, because next word can be processed in parafoveal view, and 3) there is a minimum fixation time required for acquisition of a word's meaning. Accordingly, we use fixation duration threshold of 150ms. We implemented a simple, two-state, line-oriented reading model and used it to group these lexical fixations into reading and scanning sequences (Figure 2). A reading state represents reading in one line; reading in the subsequent line is represented by a new reading state. A scanning state represents isolated lexical fixations.

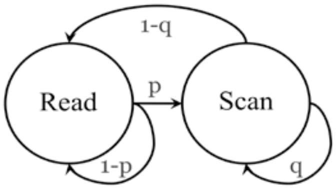

Figure 2. Two-state modeling of reading. p and q are probabilities of transitions between states

To investigate our second research question, we used reading-fixation-sequence based measures that have been suggested as reflecting aspects of cognitive effort (e.g., Rayner et al. 2006; Rayner et al. 2011).

- fixation duration,
- number of regression fixations in the reading sequence,
- the spacing of fixations in the reading sequence (referred to as perceptual span),
- reading speed, a) defined as the length of reading in pixels per unit time, b) as the number of words fixated upon per unit of time, and
- reading length, a) defined as the length of reading in pixels; b) defined as the number of words fixated upon.

We also included reaction time (RT) as a standard measure of cognitive effort. Longer reaction times indicate more effort involved in accomplishing a task. RT was defined in our study as the time from the onset of document presentation to the participant's key press expressing their relevance judgment.

We started by cleaning and pre-processing eye-tracking data. We used only those fixations where the quality of data was good and where fixation was with-in the screen coordinates. Bad quality fixations were defined as missing eye or a low probability of correct acquisition of eye position (as reported by Tobii eye-tracker). This resulted in removing approximately 5% of fixations. We considered all trials in which a participant responded by pressing yes/no, without regard to the correctness of the response. Due to the typically high individual variability of eye-tracking measures, we first personalized measures by calculating z-scores for each user separately. The underlying procedure is similar to personalization of measures described in (Buscher et al. 2012). The procedure effectively removes variability due to an individual.

We performed a series of one-way ANOVA analyses with degree of relevance as independent factor, and examined the effects of relevance on reading vs. scanning, number and duration of reading sequences, fixation duration, regressions, perceptual span and reading length and speed.

Table 1. Probabilities of transition between **reading** and **scanning** states. (S-scanning, R-reading).

Note: Due to lack of homogeneity of variance, we report Welch's corrected F. In all cases p<.001. This applies to all statistics reported in tables.

	I	T	R	ANOVA
SS	**0.52**(.013)	0.46(.013)	0.38(.013)	F(2,809)=37.6
SR	0.47(.013)	0.54(.013)	**0.61**(.013)	F(2,809)=37.2
RR	0.79(.006)	0.84(.006)	**0.88**(.004)	F(2,790)=122
RS	**0.21**(.006)	0.16(.006)	0.12(.004)	F(2,790)=122

Post hoc tests (Games-Howell) indicated that the significant differences were between all pairs of results. The highest probability of reading was for relevant documents. The highest probability of scanning was for irrelevant documents. Topical documents were in the middle (Table 2). Illustrative eye movement patterns for three selected document are shown in Figure 3.

Examining other dependent variables we found that perceptual span, the number of retrograde fixations, and the speed of reading measured in words/time were not significantly different between the document types. However, all other variables differed significantly. We report them below. Table 2 presents eye-tracking variables normalized by the length of documents (in words).

Table 2. Eye-tracking derived variables normalized by the length of documents in words.

Variable	I	T	R
reaction time (RT)	43	71	59
reading length	8.6	16	13
reading speed in px/time	.18	.21	.21
duration of reading	26	47	41
duration of scanning	13	15	9
number of reading sequences	.11	.2	.17
number of fixations on words	25	42	34
total number of fixations	.17	.27	.22

Similarly as reported for Table 1, the post hoc tests (Games-Howell) for results presented in Table 2 indicate that the significant differences were between almost all pairs of results. One exception was the lack of significant difference in reading speed in pixels between topical and relevant documents. Reaction time (normalized to the length of documents in words; Table 2) shows that judging topical documents was most effortful, while judging the irrelevant documents was the easiest. Eye-tracking based measures show a similar pattern (Table 2 and Figure 3), with an exception of reading speed, which indicated that reading irrelevant documents was slower, and reading topical or relevant documents was faster.

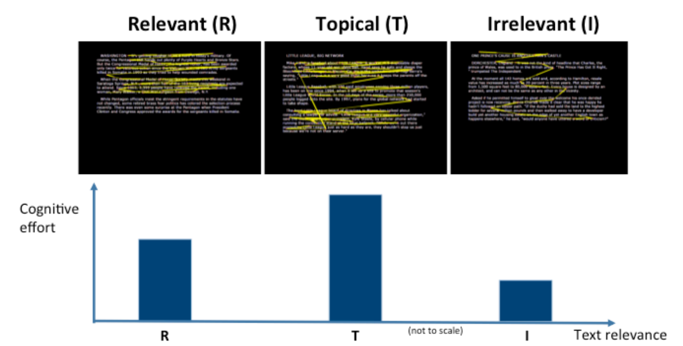

Figure 3. Example eye movement patterns and relationship between text relevance and cognitive effort.

The results obtained thus far show that the degree of relevance of a text document does affect how it is read. Significant differences in reading patterns were found between documents at the three levels of relevance. These findings generally agree with Buscher at el. (2012) in that relevant documents tend to be read more coherently, whereas irrelevant documents tend to be scanned.

The degree of relevance of a text document seems to affect cognitive effort involved in reading it. Reaction time and a collection of eye-tracking based measures indicate that the lowest cognitive effort was involved in judging that a news story is not relevant. Judging topically relevant documents required highest effort, while the effort involved judging relevant documents was generally in the middle. This result agrees with Villa and Halvey (2013), who used subjective workload judgment (NASA TLX) to investigate effort involved in relevance judgment. Their results show the same direction of relationship between effort and judging irrelevant and topical documents. However, they did not find a significant differences in cognitive effort between judging irrelevant and relevant documents. This difference in comparison with our results could be because we used a different and a more sensitive assessment of cognitive effort.

Examining the absolute measures, the duration of the longest reading sequences and the longest fixation in reading sequences (Table 3), we found that they were longest for the relevant documents. This is likely an indication that maximum of cognitive demands (Xie & Salvendy, 2001) were imposed by reading some parts of a relevant document, but that, on the average, the effort involved in processing these documents was lower than involved in processing topical documents. Overall, our results demonstrate that the degree of relevance of a text document does affect how it is read and that it does affect the level of cognitive effort required to read documents. The results indicate that relevant documents tended to be continuously read, while irrelevant documents tended to be scanned. In most cases, cognitive effort inferred from eye-tracking data was highest for partially relevant documents and lowest for irrelevant documents.

Our results largely agree with prior findings. However, our contribution is not just in confirming prior results, but also in extending them to documents with three levels of relevance and to a wider range of information topics. The latter is a likely indication that the relationships are independent of topics.

In this paper, we have reported statistical differences in reading patterns and in cognitive effort between documents of different degrees of relevance. In the follow up work, we will attempt to use our data to classify document relevance. One particularly interesting question will be to examine how much eye-tracking data on a given document is needed to plausibly classify the document's degree of relevance. An information retrieval system that knows perceived document relevance can use this information as implicit relevance feedback (White & Kelly, 2006) and return a document set that closer matches a user's search intent. Future work will also examine eye-tracking measures in relation to correctness of user relevance judgments and will look at dynamic changes of cognitive effort while a user is reading one document, before and after she encounters the relevant words. We will also check what words in relevant documents imposed the highest peak cognitive load demands.

Experiment presented in this paper complements our work on information relevance that employs fMRI (Gwizdka, 2013). At a theoretical level, our fMRI work contributes to better understanding of the multi-dimensional concept of information relevance in terms of investigating what networks of brain activations are associated with relevance judgements. At an applied level, findings from our eye-tracking experiment indicate a possibility of inferring degree of relevance of documents in real-time. We believe that establishing distinctions in brain activity related to information search should lead to a better understanding of the search process and, in combination with more applied eye-tracking approaches, to the design of better search engines.

This work was in part supported by IMLS award #RE-04-11-0062-11 and a Google Faculty Award.

References

- ❖ Borlund, P. (2003). The concept of relevance in IR. *JASIST*, 54(10), 913–925.
- ❖ Buscher, G., Dengel, A., Biedert, R., & Elst, L. V. (2012). Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM TOIS*, 1(2), 9:1–9:30.
- ❖ Cole, M., Gwizdka, J., Liu C., Bierig, R., Belkin, N., Zhang, X. (2011a). Task and User Effects on Reading Patterns in Information Search. Interacting with Computers. 23(4), 346-362.
- ❖ Cole, M., Gwizdka, J., Liu, C., Belkin, NJ. (2011b). Dynamic Assessment of Information Acquisition Effort During Interactive Search. In proceedings of the 74th Annual Meeting of the American Society for Information Science & Technology (ASIS&T 2011).
- ❖ Graff, D. (2002). The AQUAINT Corpus of English News Text, Linguistic Data Consortium, Philadelphia.
- ❖ Gwizdka, J. (2013). Looking for Relevance In Bran. Paper presented at the Gmunden Retreat on NeuroIS'2013. June 1-4, 2013. Gmunden, Austria. **Dr. H. Zemlicka most visionary paper award.**
- ❖ Rayner, K., Pollatsek, A., Ashby, J., & Jr, C. C. (2011). *Psychology of Reading*. Psychology Press
- ❖ Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, 10, 241–255.
- ❖ Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, 7, 4–22.
- ❖ Saracevic, T. (1975) . Relevance: A review of and a framework for the thinking on the notion in information science. *Society for Information Science*, 26, 321–343.
- ❖ Villa, R., & Halvey, M. (2013). Is relevance hard work?: evaluating the effort of making relevant assessments. *SIGIR'13*. (pp. 765–768). New York, NY: ACM.
- ❖ Voorhees, E.M., Dang, H.T. (2005) Overview Of The Trec 2005 Question Answering Track. Nist.
- ❖ White, R., & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. *CIKM'06*, (pp. 297-306).
- ❖ Xie, B., & Salvendy, G. (2000). Prediction of Metal Workload in Single and Multiple Task Environments. *International Journal of Cognitive Ergonomics*. 4(3), 213-242.