News Stories Relevance Effects on Eye-Movements

Jacek Gwizdka
School of Information
University of Texas, Austin, USA

Abstract

Relevance is a fundamental concept in information retrieval. We consider relevance from the user’s perspective and ask if the degree of relevance can be inferred from eye-tracking data and if it is related to the cognitive effort involved in relevance judgments. To this end we conducted a study, in which participants were asked to find information in screen-long text documents containing news stories. Each participant responded to fourteen trials consisting of an information question followed by three documents each at a different level of relevance (irrelevant, partially relevant, and relevant). 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.

CR Categories: H.3.3 Information Search and Retrieval

Keywords: eye-tracking, relevance, reading, cognitive effort

1 Introduction

The main goal of users engaged in information search is to find relevant information. Relevance, a fundamental concept in information retrieval (IR), can be considered from two perspectives system and human [Borlund 2003]. Of our interest is the second kind of relevance that has been conceptualized as the user’s judgment of the strength of the relationship between a document and the user’s information need [Saracevic 1975]. Relevance judgments can indicate user’s interest and progress in a information search task, and reflect the search system’s effectiveness. In IR studies relevance is often measured in terms of explicit user actions, such as display time or saving a document, or through self-assessments. User actions, however, provide ambiguous evidence [Kelly&Belkin 2004], while subjective assessments can be biased. Direct, non-intrusive detection of relevance judgments would provide an objective means to capture this important aspect of the user’s mental state while in the ‘flow’ of search and so enable study of search behavior in natural settings. In this work we ask two questions: 1) Does the degree of relevance of a text document affect how it is read?, and 2) Does the degree of relevance of a text document affect cognitive effort invested in reading it?

2 Related Work

A number of researchers employed eye-tracking in IR contexts with a focus on inferring relevance and on incorporating this information to improve retrieval results. Ajanki et al. [2009] used eye-movement based features as implicit relevance feedback. Eye-movement features that significantly contributed to their model included: regressions from following words and relative duration of the first fixation on a word. They showed a modest improvement in mean average precision when the eye-based features were used to select additional query terms. Buscher et al. [2012] reported on two studies in which they examined the relationship between several eye movement measures and relevance of read text passages. One search topic was examined in the first study and two topics in the second. They found that the most expressive measure with respect to relevance is based on the length of coherently read text. Surprisingly, they found fixation duration not to be a good discriminator between relevant and not relevant text. Their results showed that eye-gaze based techniques improve performance of an information retrieval ranking algorithm overall by about 8% and by about 27% for poorly performing queries. Balatsoukas & Ruthven [2012] describe an extensive study of relevance and eye-tracking measures. They showed that users expend more cognitive efforts (more frequent and longer fixations) on non-relevant document surrogates. One limitation of their study was the use of only a limited set of features (fixation duration and a number of fixations) and a cut-off of fixations shorter than 200ms, as well as talk-aloud method that might have affected length of fixations. Loboda et al. [2011] examined relationship between eye movements and word-level relevance. Only one search topic was used. They found that relevant sentence-terminal words received significantly more fixations than non-relevant words.

Our work differs in the use of text documents of varying degrees of relevance, in the use of a larger number of search topics, and in a controlled presentation of documents associated with each task. Furthermore, we use only fixations that are part of reading sequences and eliminate single lexical fixations.

3 Method

3.1 Experimental Design and Procedure

We conducted an eye-tracking experiment with N=24 participants. The experiment was conducted using Tobii T-60 eye-tracker in a usability lab. The essential experimental design is shown in Figure 1. Each subject performed a simulated information search. The task involved finding relevant factual information in news stories and was expected to require higher-level semantic processing. 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. Fourteen questions were presented in pseudo-randomized order, each followed by three texts 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 3). 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 forty-two news stories on a binary scale (yes/no). Before the actual task began, participants performed a few training trials.

### 3.2 Dataset

The documents used in the study were news stories from the AQUAINT corpus [Graff 2002]. The news stories came 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 three-level relevance assessments for the documents from TreC Q&A task from 2005 [Voorhees & Dang 2005].

We processed the news stories into experimental stimuli using the following procedure. First, we created HTML documents from the new stories to automate capture of AOIs for each word as it was rendered on screen. Second, we marked up relevant words (i.e. words that contained an answer to the question) in relevant documents and stored these words along with their coordinates in a database. Third, we took a screen shot of each HTML document (without any markup of the relevant words) and used the resulting images as text stimuli (screen resolution was kept the same).

### 3.3 Eye movement analysis

We use a similar approach to [Cole et al., 2011] which is briefly summarized below. We implemented a simple, two-state, line-oriented reading model incorporating main assumptions from the E-Z Reader model [Reichle et al, 2006; Rayner et al, 2011], such as that lexical processing of words is serial, one word at a time in the order of word appearance in text, and that more than one word can be processed on a fixation, because next word can be processed in parareal view. We used a minimum fixation duration threshold of 150ms to select eye fixations that were likely to result in word meaning acquisition. We used our algorithm to process the location and duration of participant eye fixations as captured by eye-tracker. Our reading model was then used 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.

### 3.4 Measures of Cognitive Effort

Several cognitive effort measures based on reading fixation sequences have been suggested [Rayner et al. 2006; Rayner et al. 2011]. We use the following measures:

- fixation duration,
- number of regression fixations in the reading sequence,
- the spacing of fixations in the reading sequence (we refer to it as perceptual span),
- reading speed, defined as the length of text acquired per unit time, and
- reading length, defined as the length of reading in pixels.

We also included reaction time (RT) as a standard measure of cognitive effort. Longer reaction times indicate more effort involved in accomplishing a task.

### 4 Results

We consider all trials in which a participant responded by pressing yes/no, without regard to the correctness of the response. We cleaned the eye-tracking data and used only those fixations where the quality of data was good and where fixation was within the screen coordinates. Bad quality fixations were defined as missing eye (reported by Tobii eye-tracker as validity=4) or a low probability of correct acquisition of eye position (validity=3). This resulted in removing approximately 5% of fixations (out of the total of 76,778 fixations). The number of trials that were left was not equally distributed between the three relevance levels, I/T/R: 333/235/298. Additionally, irrelevant documents tended to be longer. To obtain a more uniform distribution of document relevance degrees and their text lengths, we removed data from irrelevant documents longer or equal 310 words. After this operation was completed, we obtained the following distribution of document relevance levels, I/T/R: 283/235/294. The difference was now around 20%, which is acceptable, given robustness of ANOVA. At the same time, the average length of the texts left in the analysis pool was for I/T/R respectively: 178/177/178 (SD=30).

![Figure 2](image2.png)

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

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, and thus it removes relatedness of the measurements.

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.
Games u-h-n-c-s. F-n-n-c-i- are in the i-, who used b-ing it. Reaction time and eye d-c-or a p-pr contribution is a-e-u-ants), we analyzed data for each participant sep-

Table 1. Probabilities of transition between reading and scanning states (mean(stderr)). Refer to Figure 2 for explanation of transitions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>I</th>
<th>T</th>
<th>R</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>q (SS)</td>
<td>0.51(0.14)</td>
<td>0.46(0.15)</td>
<td>0.38(0.15)</td>
<td>F(2,519)=27</td>
</tr>
<tr>
<td>1-q (SR)</td>
<td>0.49(0.14)</td>
<td>0.53(0.15)</td>
<td>0.61(0.15)</td>
<td>F(2,519)=27</td>
</tr>
<tr>
<td>1-p (RR)</td>
<td>0.78(0.07)</td>
<td>0.84(0.07)</td>
<td>0.88(0.066)</td>
<td>F(2,506)=91</td>
</tr>
<tr>
<td>p (RS)</td>
<td>0.22(0.07)</td>
<td>0.16(0.07)</td>
<td>0.12(0.066)</td>
<td>F(2,515)=88</td>
</tr>
</tbody>
</table>

Examining other dependent variables we found that perceptual span and the number of retrograde fixations were not significantly different between the conditions. However, all other variables differed significantly. We report them below. Table 2 contains variables normalized by the length of documents, while Table 3 contains absolute variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>I</th>
<th>T</th>
<th>R</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>reaction time (RT)</td>
<td>L</td>
<td>H</td>
<td>M</td>
<td>F(2,517)=119</td>
</tr>
<tr>
<td>reading length</td>
<td>L</td>
<td>H</td>
<td>M</td>
<td>F(2,511)=93</td>
</tr>
<tr>
<td>reading speed</td>
<td>L</td>
<td>H</td>
<td>F</td>
<td>F(2,529)=22</td>
</tr>
<tr>
<td>duration of reading</td>
<td>L</td>
<td>M</td>
<td>H</td>
<td>F(2,508)=100</td>
</tr>
<tr>
<td>duration of scanning</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>F(2,517)=45</td>
</tr>
<tr>
<td>number of reading sequences</td>
<td>L</td>
<td>M</td>
<td>H</td>
<td>F(2,517)=76</td>
</tr>
<tr>
<td>total number of fixations</td>
<td>L</td>
<td>M</td>
<td>H</td>
<td>F(2,512)=96</td>
</tr>
</tbody>
</table>

Table 2. Eye-tracking derived variables normalized by the length of documents in words. L-indicate lowest value, M-middle value, H- highest value. Two H in one row indicate no significant difference between the document relevance conditions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>I</th>
<th>T</th>
<th>R</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>duration of a longest reading seq</td>
<td>L</td>
<td>M</td>
<td>H</td>
<td>F(2,497)=80</td>
</tr>
<tr>
<td>longest fixation in a reading seq</td>
<td>L</td>
<td>M</td>
<td>H</td>
<td>F(2,518)=35</td>
</tr>
</tbody>
</table>

Table 3. Absolute eye-tracking based variables.

Post hoc tests (Games-Howell) indicate that the significant differences were between all pairs of results (typically p<.001). The highest probability of reading was for relevant documents, the lowest for irrelevant documents. Topical documents were in the middle. In addition to the analysis across all participants presented in Table 1, we analyzed data for each participant separately. This analysis showed the same trend – in 71% to 92% cases the highest probability of transition agreed with Table 1. On the average, 52% of cases where the patterns agreed were significant, which is better than the chance probability of 33%. The relationships were strongest for probability of transitions from reading state, where 63%-67% cases were significant.

Examining other dependent variables we found that perceptual span and the number of retrograde fixations were not significantly different between the conditions. However, all other variables differed significantly. We report them below. Table 2 contains variables normalized by the length of documents, while Table 3 contains absolute variables.

Similarly as reported for Table 1, the post hoc tests (Games-Howell) for results presented in Table 2 and 3 indicate that the significant differences were between almost all pairs of results (typically p<.001). One exception was lack of significant difference in reading speed between topical and relevant documents. Our separate analyses at the level of individual participants showed in most cases (for 80-96% of participants) and for most normalized variables, patterns similar to the results of analysis across all participants presented in Table 2. For a duration of scanning the agreement was somewhat lower (for 71% of participants). In all cases, the number of individual participants for whom the patterns were confirmed, is larger than the chance probability.

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.

Normalized reaction time (i.e. time to from the onset of text stimulus presentation to the participant’s key press expressing its relevance judgment) is a simple measure of cognitive effort and shows that the judging topical documents was more effortful, while judging the irrelevant documents was the easiest. Eye-tracking based measures show a similar pattern, with an exception of reading speed, which indicates that reading irrelevant documents was slower, and reading topical or relevant documents was faster.

5 Discussion

Revisiting our two research questions. The results show that the degree of relevance of a text document does affect how it is read. We found significant differences in reading patterns between documents at the three levels of relevance. Our findings generally agree with Buscher et al. [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 eye-tracking based measures (including: number of reading sequences, duration of reading fixations, total number of fixations, the length of text read – all normalized by document length) indicate that the lowest cognitive effort was involved in judging that a news story is not relevant to the question. Judging topicaly relevant documents required highest effort, while the effort involved judging relevant documents was generally in the middle. This finding agrees overall with Villa and Halvey [2013], who used subjective workload judgment (NASA TLX) to investigate effort involved in relevance judgment (using texts from the same document set). Their results showed the same direction of relationship between effort and judging irrelevant and topical documents. However, their results did not find a significant differences in cognitive effort between judging irrelevant and relevant documents. This difference could be because of we employed a different measurements of cognitive effort. Our results differ from Balatsoukas & Ruthven [2012], who found longest fixation durations involved in judging irrelevant document surrogates. That difference is likely to indicate differences in reading documents and their surrogates (e.g., search results snippets displayed by a search engine).

The differences in reading patterns and in cognitive effort obtained at the aggregate level of all participants, were confirmed to a large degree in separate analyses at the level of individual participants. This makes application of eye-tracking in the method of inferring document relevance quite promising.

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 a 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.

6 Conclusions

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 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 rela-
tionships are independent of topics. This conclusion is corrobo-
rated by results from ANOVA with question topic as a factor.
While the topic effect was overall significant, the post-hoc tests
showed that this effect was, in most cases, due to just one topic.

One study limitation lies in our construction of reading states.
Continuous reading in each line is considered a separate reading
state. This limitation, however, does not impact the measures
used in this paper.

In this paper, we tested hypotheses of 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 of
eye-tracking data on a given document is needed to plausibly
classify the document’s degree of relevance. An information
retrieval systems 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.

7 Acknowledgments

This work was in part supported by IMLS award #RE-04-11-
0062-11 and a Google Faculty Award. Thanks to Michael Cole
for great help with this project.

References

Ajani, A., Hardoon, D., Kaski, S., Puolamäki, K., & Shawe-
from gaze patterns. User Modeling and User-Adapted
Interaction, 19(4), 307–339

to the analysis of relevance judgments on the Web: The case of
Google search engine. Journal of the American Society for
Information Science & Technology, 63(9), 1728–1746.

Borlund, P. (2003). The concept of relevance in IR. Journal of
American Society for Information Science & Technology,
54(10), 913–925.

Attentive documents: Eye tracking as implicit feedback for
information retrieval and beyond. ACM Transactions on
Interactive and Intelligent Systems, 1(2), 9:1–9:30.

Cole, M., Gwizdka, J., Liu C., Bierig, R., Belkin, N., Zhang,
X. (2011). Task and User Effects on Reading Patterns
in Information Search. Interacting with Computers. 23(4),
346–362.

Text, Linguistic Data Consortium, Philadelphia.

Kelly, D. and Belkin, N.J. (2004). Display Time As Implicit

Inferring word relevance from eye-movements of readers. In

for the thinking on the notion in information science. Society
for Information Science, 26, 321–343.

Psychology of Reading. Psychology Press

Eye movements as reflections of comprehension processes in
reading. Scientific Studies of Reading, 10, 241–255.

Reader: A cognitive-control, serial-attention model of eye-
movement behavior during reading. Cognitive Systems
Research, 7, 4–22.

evaluating the effort of making relevant assessments. In

2005 QUESTION ANSWERING TRACK. NIST.

personalization and task information on implicit feedback

in Single and Multiple Task Environments. International

Does the next news story contain the following information:
Russian submarine Kursk sinks: Which Russian fleet was the submarine part of?

Figure 3 Example of a question and news text stimuli in one trial consisting of three stories at varying levels of relevance.