Attentive Documents: Eye Tracking as Implicit Feedback for Information Retrieval and Beyond

GEORG BUSCHER, ANDREAS DENGEL, RALF BIEDERT, and LUDGER VAN ELST,
German Research Center for Artificial Intelligence (DFKI)

9

Reading is one of the most frequent activities of knowledge workers. Eye tracking can provide information on what document parts users read, and how they were read. This article aims at generating implicit relevance feedback from eye movements that can be used for information retrieval personalization and further applications.

We report the findings from two studies which examine the relation between several eye movement measures and user-perceived relevance of read text passages. The results show that the measures are generally noisy, but after personalizing them we find clear relations between the measures and relevance. In addition, the second study demonstrates the effect of using reading behavior as implicit relevance feedback for personalizing search. The results indicate that gaze-based feedback is very useful and can greatly improve the quality of Web search. The article concludes with an outlook introducing attentive documents keeping track of how users consume them. Based on eye movement feedback, we describe a number of possible applications to make working with documents more effective.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance feedback*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms: Experimentation, Human Factors, Measurement

Additional Key Words and Phrases: Relevance feedback, eye movement measures, personalization, attentive documents

ACM Reference Format:
Buscher, G., Dengel, A., Biedert, R., and Van Elst, L. 2012. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. ACM Trans. Interact. Intell. Syst. 1, 2, Article 9 (January 2012), 30 pages.
DOI = 10.1145/2070719.2070722 http://doi.acm.org/10.1145/2070719.2070722

1. INTRODUCTION

Forty years ago, Herbert Simon pointed out that *attention* is a scarce resource in an information-rich world [Simon 1969, 1971]. Therefore, the way we do utilize our "attention budget" reflects to a certain degree our own preferences: We try to focus on these pieces of information that we think are most relevant, interesting, or useful to us in our current situation. Obviously, in an attention economy, where attention might be understood as a currency [Davenport and Beck 2001], there may exist many factors

This work was supported by the German Federal Ministry of Education, Science, Research, and Technology (bmb+f): Project Mymory, grant 011WF01; Project Perspecting, grant 011W08002.
Authors' addresses: G. Buscher (contact author), Microsoft Corporation, One Microsoft Way, Redmond, WA 98052; email: georg@gbuscher.com; A. Dengel, R. Biedert, L. Van Elst, Knowledge Management Department, German Research Center for Artificial Intelligence (DFKI), Trippstadter Strasse 122, 67663, Kaiserslautern, Germany.

DOI 10.1145/2070719.2070722 http://doi.acm.org/10.1145/2070719.2070722

that distract us from a self-directed utilization of our "attention budget." The tight coupling of Web search and advertising is just one example. Either way, the extent to which a piece of information is interesting or relevant to us influences how we pay attention to it and how we perceive it.

This is especially true for textual information, which, in its various manifestations, has become an important source of knowledge in professional as well as in private settings. For example, we read emails, reports, papers, blogs, presentation slides, etc., in order to acquire information, establish our opinions, and base our decisions. *What* we read is influencing what we think and what we know; and *how* we read reflects, amongst other cognitive dispositions, the handling of our attention resources. For example, on the one hand, we may skim parts of a document being of little interest or of which the content is already known to us. On the other hand, we may intensively read text if it seems relevant or interesting. In some cases, we may get stuck or ignore text that is too difficult to understand or too detailed for us to follow.

While we do not have direct access to attention and cognitive activity during reading, there is evidence that eye movements are tightly coupled with these processes in our brains [Liversedge and Findlay 2000] and may serve as measurable indicators for such processes. Therefore, by analyzing eye movements in detail, we might be able to infer how the reader perceived a text he or she has read, that is, whether it was relevant, irrelevant, too difficult to understand, etc.

Technology for measuring eye movements has evolved rapidly during the last couple of years. Nowadays, eye tracking devices are relatively unobtrusive to use and have sufficiently high precision to be useful for analyzing eye movements during reading. Whereas good commercial eye trackers are still too expensive to be widely employed at office workplaces, a boost in the development of low-fidelity, open-source eye-tracking systems and software could be observed.[1] It is conceivable that they will further develop so that eye tracking systems will get more wide-spread in the future.

All in all, affordable eye tracking technology together with methods to interpret that data in terms of a reader's attention not only gives us the opportunity to analyze how humans deal with that scarce resource. It also has the potential to facilitate the development of information systems that are optimized in this regard and to design new kinds of applications. For example, documents may become *attentive* in the sense that they record the way they are consumed by the reader. Such usage information could be stored in association with the documents so that the documents do not forget when, how, and what parts of them the reader paid attention to. These attentive documents could, for example, form the basis for improving and personalizing search, making the reading process more effective, and supporting collaborative work.

The purpose of this article is to explore how useful eye tracking is as an interactive method to generate implicit feedback while reading. Within the scope of this article we focus on interpreting and using eye movements as implicit relevance feedback in Web search scenarios. We analyze eye gaze data in order to determine what parts of a document were read, and which of these parts were also relevant to the reader. Therefore, we implemented an algorithm that analyzes raw gaze data from the eye tracker and detects reading behavior. Based on eye movements during reading, we compute several gaze-based measures and explore their relation to individually perceived relevance of read text in two eye tracking studies. We show that the amount of reading and skimming changes with relevance and topicality of the viewed text and demonstrate that one of the most expressive measures for relevance is coherently read text length, that is, the length of text the user has read line by line without skipping any part.

[1]http://www.gazegroup.org/.

We then apply eye tracking together with the studied measures of relevance in order to determine *which parts* of a text the reader has read, *how intensely* these parts have been read, and finally to estimate *how relevant* they have been to him or her. Finally, we demonstrate the usefulness of the reading detection algorithm and the studied eye gaze measures in the application domain of Web search. Here we show that implicit relevance feedback from eye tracking can greatly improve the quality search engine result rankings by over 25% compared to the original ranking, and by over 8% compared to baseline reranking methods based on nongaze implicit feedback.

The remainder of the article is structured as follows: First, we give an overview of related research (Section 2) and explain the foundations for our basic reading detection method (Section 3). We then explore the relation of several eye movement measures to user-perceived relevance of read text. To this end, we report on the results of two eye tracking studies employing two different reading scenarios (Sections 4 and 5). We demonstrate the value of gaze-based implicit relevance feedback for personalization in information retrieval (Section 5), and finally give an outlook of further applications for attentive documents that seem possible in the future.

2. BACKGROUND AND RELATED WORK

In the human-computer interaction domain, eye tracking has been applied primarily in diagnostic applications to learn how interfaces are used [Cutrell and Guan 2007; Buscher et al. 2009a] and directly in interactive applications such as eye typing [Majaranta and Räihä 2002].

In our work, we aim at interactive gaze-based applications, however, not based on direct control from the eyes. Since the eyes naturally did not evolve as tools but rather as perceptual organs and since many eye movements are unconscious, it is often difficult to use gaze for direct control. However, since eye movements are relatively tightly coupled with cognitive processes [Liversedge and Findlay 2000], there is a great deal of implicit information that may be inferred by analyzing eye movements in our daily work. This implicit information, for instance, about how and where we paid attention is very valuable feedback that can be used to support the individual user's work. Therefore, we aim at using gaze information in a passive, indiscernible way so that the computer system can assist the user.

In the following, we review related work first with respect to reading-related eye gaze feedback. Then we particularly focus on the application of implicit feedback in the information retrieval domain, how implicit feedback is commonly assessed, and how eye tracking could be of use here.

2.1. Towards Attentive Documents

There is only very limited research focusing on the broader idea of what we call attentive documents, that is, documents actively keeping track of how they are consumed by users and applications based on this data.

One of the first occurrences of a similar idea can be found in a paper by Hill et al. [1992]. They introduce the interesting concept of computational wear for digital documents in the style of wear for physical documents. Documents actively keep track of user activities with respect to reading (i.e., read wear) and editing (i.e., edit wear). This is done by measuring display time of different document parts and by recording editing interactions. In their implemented version, wear (i.e., the amount of usage of document parts) is visualized on the scrollbars for a document. The authors argue that such information can be very helpful for professional document work. For example, read wear could be used to show the user which parts of a document have been most interesting to him/her before. In that way, it can provide support for improved refinding and recognition. Edit wear could be useful to get informed about document parts that

have been edited by coauthors. The latter concept can now be found in most modern word processing software applications.

Ohno [2004] develops the concept of read wear further and applies an eye tracker to keep track of eye gaze traces on documents. The central idea is to improve document browsing by making it easier to find and recognize information read before in long documents. The main conceptual advancement of our method over the system implemented by Hill et al. is the employment of measures derived from gaze traces in order to determine the intensity of reading. Therefore, the author uses fixation duration on document parts as a direct and simple measure for the approximation of reading intensity.

Xu et al. [2009] also apply eye tracking to record how much a user pays attention to different document parts. Similar to Ohno, they use fixation duration on document terms as a direct measure of reading intensity (or "interestingness"). However, in contrast to Ohno, they use this kind of information as implicit feedback in order to generate personalized document summaries. Such personalized summaries emphasize these parts and terms of a document that have been read intensively before. The authors show that summaries generated in this way are more reflective of a user's reading interest and summary preference.

In a more interactive way, Hyrskykari et al. [2003] analyze eye movements on the fly while a user is reading and tries to automatically detect comprehension difficulties on foreign language texts. The basic idea is to use the total gaze duration on a word in order to decide whether help for this word or the entire sentence should be provided. Overall, the work shows that it is difficult to detect understanding difficulties on the fly. However, the general idea is intriguing since the text being read plays an active role in the reading process because it actively observes and helps the reader understand it.

2.2. Implicit Feedback for Information Retrieval

Implicit relevance feedback for information retrieval is an important area of research and has been tackled by many different researchers (see Kelly and Teevan [2003] for an early overview). It has been shown that incorporating explicit relevance feedback, that is, asking the user directly for relevance judgments of documents for a given query, can drastically improve the quality of search result rankings [Rocchio 1971; Salton and Buckley 1990]. However, explicitly asking users for relevance judgments imposes an additional burden on the user and causes additional cognitive load. Therefore, work about implicit relevance feedback generally aims at analyzing user context or interpreting user interactions in order to generate relevance feedback for search implicitly, without causing any additional cognitive load for the user. Most of the research in this area is based on implicit feedback from user context or user interactions not related to eye tracking. However, there is also some recent work based on gaze data.

2.2.1. Non-Gaze-Based Feedback. One of the most frequently researched implicit feedback data sources is display time, that is, the duration a document is visible on the screen. As one of the earliest, Morita and Shinoda [1994] analyzed reading time for documents and its correlation to explicit relevance judgments for these documents. They found a strong tendency of users to spend more time on interesting documents than on uninteresting ones; a finding that has also been reported by Claypool et al. [2001] and Fox et al. [2005]. However, in more complex and naturalistic settings, Kelly and Belkin could not find any general relationship between display time and the users' explicit relevance judgments for documents [Kelly and Belkin 2001, 2004. But when taking different task types into account White and Kelly [2006] found clear signals in display time and showed that retrieval performance could be improved. Buscher et al. [2009b] analyze display time in greater detail and measure the duration different document

parts have been visible on the screen. They report considerable improvements when employing such fine-grained implicit feedback for the improvement of search result ranking. In addition, Gyllstrom [2009] shows that re-finding information in documents on the desktop can be sped up when incorporating information about which document parts had been visible on the screen before.

In addition to display time for documents, it has been found that the amount of scrolling on a Web page [Claypool et al. 2001], click-through for documents in a browser [Fox et al. 2005; Joachims et al. 2007], and exit type for a Web page [Fox et al. 2005] are good indicators of interest. Agichtein et al. [2006] have shown that search result ranking can be considerably improved when combining implicit feedback from several implicit relevance indicators. In important finding by Melucci and White [2007] is that individually personalizing such measures leads to additional improvements.

Most of the previously cited work was based on implicit feedback acquired for entire documents. However, there is only little work focusing on implicit feedback on the segment level, that is, implicit feedback for document *parts*. [Golovchinsky et al. 1999] recorded highlighting and underlining behavior of users working with documents and also analyzed notes in margin of a document in order to infer individually perceived relevance of different document parts. Using this kind of implicit feedback, the authors showed that the quality of search can be significantly improved. A somewhat similar approach has been examined in Ahn et al. [2008] where users could copy and paste interesting document parts in a personal notebook. They also reported improvements of search engine ranking.

2.2.2. Gaze-Based Feedback. There is some research focusing on gaining implicit relevance feedback from eye movements. Balatsoukas and Ruthven [2010] compute basic eye movement measures from eye movements on search engine results pages and investigate their expressiveness with regard to a variety of relevance criteria. They report that result entries that differ in topicality have strongest effects on eye movement measures, but interestingly criteria like familiarity with the information content also play a role. Moe et al. [2007] use a qualitative and exploratory approach and identify eye movement measures that are correlated with explicit relevance ratings for the read text. While most of their tested measures were inconclusive, they found that the amount of reading behavior is informative with respect to relevance of the read text. Loboda et al. [2011] inspected the signal in eye movement measures for inferring relevance of single words and found that sentence-terminal words in particular attract more and longer fixations. Brooks et al. [2006] also focused on determining what gaze-based measures are most helpful in estimating relevance of single paragraphs in documents. In a preliminary study, they found that relevant text parts caused a higher number of fixations and regressions. Furthermore, Puolamäki et al. [2005] combined a number of eye tracking measures and trained HMM-based classifiers to predict relevance for previously read text. They found that predicting relevance by analyzing eye movements is possible to some extent. However, they did not examine which gaze-based measures were most correlated with relevance. Ajanki et al. [2009] further built on this work and aimed at automatically inferring queries from eye movements while reading. They trained support vector machines based on gaze data to determine text parts that have been relevant to the user.

Buscher et al. [2008b] used eye movement measures to detect parts of longer documents that have been read intensively. They used these read parts of documents as implicit feedback for query expansion and re-ranking in an information retrieval scenario and reported considerable improvements of search engine result quality as measured by explicit user ratings. Cole et al. [2010] further investigated the effect

of different tasks on reading behavior and found that switches between reading and skimming behavior are implicit indicators of the current task.

In summary, gaze is an excellent data source for providing information about how much attention the user paid to what locations and contents on the screen. Therefore, it seems to be well suited to provide information about which documents and document parts have been interesting or relevant to the user. As previous literature already indicates to some extent, this knowledge can be very valuable as implicit relevance feedback for information retrieval applications.

In this article, we examine the suitability of different eye movement measures for inferring relevance feedback *on the passage level* in detail. First, we review findings from research in reading psychology. Motivated by these findings, we introduce a reading detection method from Buscher et al. [2008a] and describe how we compute a variety of different eye movement measures. Next, we report results from two eye tracking user studies aiming at finding most useful features for detecting relevance. Finally, to show the applicability of the results from the first two studies, we report on extended analysis of a study previously partly published in Buscher et al. [2009b] demonstrating the value of gaze-based implicit relevance feedback in an information retrieval scenario.

3. EYE TRACKING

Interpreting eye tracking data can be tricky and requires careful consideration. Generally, eye movements are relatively tightly coupled with cognitive processes [Liversedge and Findlay 2000]. Therefore, they contain valuable information making it possible to infer information about processes in the brain. However, it has to be born in mind that eye movements are often unconscious in nature and there are likely to be a large number of unknown factors influencing them. Hence, gaze data is generally very noisy. In addition, eye trackers introduce further inaccuracy when estimating the eyes' focal point.

Most previous research employs gaze-based measures in a rather simple and direct way. For example, a fixation on a word is usually interpreted as user interest in or relevance of that word. However, gaze data is noisy and fixations can have different causes depending on the cognitive state of the user. Therefore, a fixation on a word may not always be meaningful. To reduce noise in gaze data, our approach is to detect reading behavior first based on specific spatial and temporal patterns in eye gaze traces. Then, we only focus on gaze data during reading behavior and ignore everything else. While in general, the focal points of our eyes are not always related to the point of visual attention, there is strong evidence that they match during reading behavior [Rayner 1998]. Hence, the gaze data we are relying on for our analysis is recorded only while users are reading, that is, while they are actively consuming and thinking about textual contents.

3.1. Reading Psychology

Automatically detecting reading behavior by analyzing eye gaze patterns is one of the key steps we take to reduce noise from eye trackers. Implementing a robust reading detection method is possible when taking into account existing knowledge from the area of reading psychology. Here, a great deal of research has been done during last one hundred years concerning eye movements while reading. When reading silently, as summed up in Rayner [1998], the eye shows a very characteristic behavior composed of fixations and saccades. A fixation is a time of about 250ms on average during which the eye is steadily gazing at one point. A saccade is a rapid, ballistic eye movement from one fixation to the next. A typical left-to-right saccade is 7–9 letter spaces long. Approximately 10–15% of the eye movements during reading are regressions, that is, movements to the left along the current line or to a previously read line. Visual

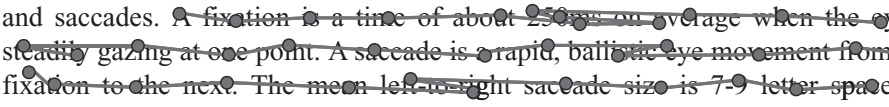

Fig. 1. Typical eye movement pattern while reading. Circles mark fixations; connecting lines depict saccades.

information can only be perceived during fixations and not during saccades. Words can be identified only up to approximately 7–8 letter spaces to the right of the fixation point. However, the total perceptual span, where at least some useful information about the text can be extracted, extends about 14–15 letter spaces to the right of the fixation point.

There is high variability of the aforementioned average values both with respect to individual differences between readers as well as with respect to document-induced differences for the same reader. For example, the fixation durations for the same reader can vary between 100–500ms, while saccade sizes can range from 1 to 15 characters. Among many other factors, this variability is influenced by the difficulty of the read text, word predictability, background knowledge, and reading strategy of the reader [Rayner 1998].

3.2. Reading Detection Method

We used these insights about typical eye movements while reading to implement a reading detection method. The method works in several steps. First, fixations are detected by grouping together nearby gaze coordinates for a duration of at least 100ms. Second, the transitions from one fixation to the next, that is, the saccades, are classified according to their distances and directions. Classification of saccades is based on a set of heuristics derived from previous research in reading psychology [Rayner 1998]. In that way, typical saccades belonging to reading behavior can be detected and differentiated from unrelated eye movements. Third, using some simple heuristics, it is determined whether a sequence of saccades over a line of text is more characteristic for reading or for skimming behavior based on the lengths and the composition of the saccades. Figure 1 shows an example for the typical placement of fixations and saccades while reading. More details about the reading detection method can be found in Buscher et al. [2008a].

3.3. Eye Movement Measures

Eventually, one of our goals for this work was to detect what parts of a document have been interesting or relevant to a user and to use this knowledge for the personalization of information retrieval methods. Most previous research applying eye tracking for information retrieval used gaze data directly to infer interestingness or relevance of single terms in a text. For example, Ohno [2004] and Xu et al. [2009] computed very simple gaze-based measures like fixation duration on every term in a document and then determined the "best," for instance, most looked at, terms for further use.

However, the contents and the relations a text conveys are mostly based on specific combinations of many words in sentences and paragraphs rather than single isolated terms. For example, regressions on a word in a text do not necessarily mean interest or relevance of that word but can rather be caused by difficulties in understanding the word or the entire sentence. However, what the regression can tell us is that the reader pays close attention to the text and tries to understand it. Therefore, in order to detect relevance of document parts, we do not apply gaze-based measures on single terms directly, but aggregate them on the level of sentences and paragraphs. Overall, we aim to determine relevance at the level of paragraphs. The basic assumption is that

if, for example, a paragraph is interesting or relevant to the reader, then this will be reflected in his or her eye movement pattern over that text part.

Hence, all of the gaze-based measures described in the following are computed as average values of fixations and saccades on entire paragraphs of text. Note, that we only include fixations and saccades belonging to reading or skimming behavior for the computation of the measures. For the following studies, we are focusing on 5 different gaze-based measures.

—*Average fixation duration* is computed as the sum of the durations of all fixations on a paragraph divided by the number of fixations on that paragraph. There is abundant evidence that fixation duration is influenced by the text is currently being fixated [Rayner 1998]. Some previous work uses this measure directly as an indicator of relevance [Ohno 2004; Xu et al. 2009]. However, it is not clear yet to what extent it relates to relevance.
—*Average forward saccade length* is the average length of left-to-right saccades. Saccade size is also known to be influenced by characteristics of the text [Rayner 1998].
—*Regression ratio* is computed as the number of regressions divided by the total number of saccades on a paragraph. There is some indication in previous research stating that higher regression ratios signify relevance of the read text [Moe et al. 2007; Brooks et al. 2006].
—*Thorough reading ratio* is computed as the length of text that has been detected as read by our reading detection method divided by the length of read or skimmed text. Thus, it is a measure for the reading intensity of a user: the more parts of a paragraph read instead of skimmed, the higher the value of this measure. A similar measure has been found to be related to relevance of read text [Moe et al. 2007].
—*Coherently read text length* measures the length of text in characters that has been read coherently without skipping any text in between. The assumption underlying this measure is that users may start and quickly stop reading a paragraph if the contained information is irrelevant. Then, they may skip the irrelevant paragraph and jump to the next one to continue. In contrast, if the information seems relevant, users may continue reading line by line without skipping any text in between.

3.4. Personalization of the Measures

As stated in Rayner [1998], there is high variability of most eye movement measures both within as well as between readers. Therefore, it is difficult to build methods estimating relevance of read text based on absolute values of gaze-based measures (e.g., compare Moe et al. [2007]). However, when individually personalizing such measures, they become more expressive so that precise implicit relevance feedback is easier to achieve as previous research suggests for non-gaze-based feedback data [Melucci and White 2007].

We personalize each of the recorded gaze-based measures as follows. First, we determine the distribution of a measure for an individual user by analyzing all of his or her recorded eye movement data during reading. Then, we compute the upper and lower whiskers (limits) concerning the measure's value distribution as it is typically done for generating box plots (i.e., lower whisker = max(minimum value, lower quartile − 1.5 * interquartile range); upper whisker = min(maximum value, upper quartile + 1.5 * interquartile range); see Figure 2) [Wilcox 2005]. The upper and lower whiskers define a user-specific interval containing most of the measured values. Outliers do not fall within such intervals and, hence, do not distort them.

Next, the computed absolute values of the eye movement measures are normalized with respect to the individual whisker-intervals. This results in a percentage for each absolute value stating its relative position in the appropriate interval. Outliers that

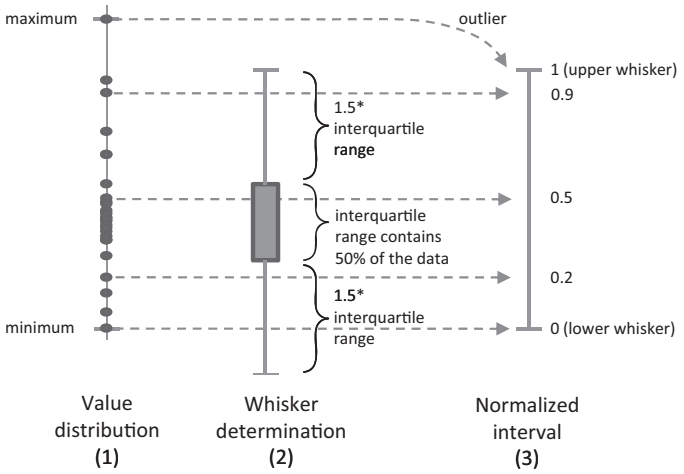

Fig. 2. Personalization of a measure by mapping absolute values into a [0,1]-interval.

do not lie within a whisker-interval are mapped to 0 or 1 depending on whether they are smaller or greater than the lower or upper whiskers, respectively. Figure 2 shows an example for a measure's value distribution, the respective determination of the whiskers, and the normalization of the measured absolute values.

4. STUDY 1: RELEVANCE ESTIMATION BASED ON EYE MOVEMENTS

The first study is of exploratory nature and analyzes eye movement measures while reading texts with varying degrees of difficulty, novelty, and topicality. Specifically, we 1. examine the relation between reading behavior and explicit user judgments, and 2. analyze the effect of measure personalization.

4.1. Methods

4.1.1. Experimental Design and Procedure. We designed an eye tracking experiment where 19 participants had to read through a number of short documents and judge their relevance with respect to a research task we provided.

The task. The same task was given to all participants. They were told to prepare a presentation about technical aspects of renewable energies. Therefore, we provided them with documents their presentation should be based on. Their task was to read through all of the documents and judge them according to their individual assessments of usefulness for the presentation.

In order to make the task more realistic, we told the participants that the experiment was composed of two different parts. The first part described above, that is, looking through and judging the usefulness of each document, was just to categorize the documents for later use. They were told that during the second part, that is, presentation writing, the documents would be presented to them again in an order based on their own categorization from the first part. However, this imaginary second part was made up only to let the task seem more realistic. An experiment has always been stopped after the judgment phase was finished.

The documents. We provided all participants with the same set of 16 documents from Wikipedia or newspapers in randomized order. 14 of the documents were on topic and dealt with different aspects of renewable energies, for instance, characteristics of renewable resources, technical explanations of machines, political opinions, and

examples of power plants. The remaining two documents were completely off topic (one article was about theory of history; one was about a famous castle).

The 14 on-topic documents were carefully selected to cover a broad range of difficulty and novelty. Some were rather easy to understand, containing only facts that could be assumed to be known by most participants, and some contained rather difficult and complicated contents, for instance, one document explained chemical and physical processes in photovoltaic solar panels on the atomic level. Also, some documents were closer and some more distantly related to the main task about technical aspects of renewable energies. For example, some documents dealt with technical explanations whereas others covered political aspects.

All of the documents were approximately one screen page long and contained about 350 words on average. A font size of 12pt was chosen. A line of text was about 720 pixels wide, which corresponded to approximately 120 characters. The documents were written in German and consisted only of plain text.

Explicit judgments. We provided the participants with the following categorization scheme for their explicit judgments.

—"Useful for the presentation" → label: "*useful*"
—"Useful, but contents completely known" → label: "*known*"
—"Probably useful, but too difficult to understand" → label: "*difficult*"
—"Rather useless" → label: "*useless*"

This categorization scheme for explicit judgments needs some discussion. Researchers in the domain of information retrieval usually employ one-dimensional relevance judgment scales with the two poles "relevant" and "irrelevant". However, there is a multitude of factors and dimensions influencing how users perceive the quality (or relevance) of a document for solving their current task. For example, as summed up by Chen and Xu [2005], relevance is influenced by topicality, novelty, reliability, understandability, and scope of the contents.

The contents of documents can be characterized in all of these dimensions independently. All of these independent characteristics can cause different behavior in the reader, and different eye movement patterns in particular. For example, if a document is perfectly on topic but already completely known to the reader, he or she may read it differently than a document that is only partly on topic but that contains many new and interesting facts. Probably the reader will come to the same personal explicit judgment for both documents in the end with respect to a one-dimensional relevance judgment scale. However, we expect to see differences in reading behavior on both documents due to the differences in novelty and understandability.

Out of the five dimensions of relevance named above, novelty and understandability are presumably most dependent on individual users. Individual characteristics like knowledge, background, vocabulary, and intelligence influence both individually perceived novelty as well as understandability. Since we expected to see differences with respect to both dimensions in the recorded eye movements, we decided against the typical one-dimensional judgment scale but decided for the four point categorization scheme including the two entries "known" and "difficult." Now, the differences between the four categories have to be clarified: The category "useful" is used for documents that are relevant on all (or most) of the five dimensions of relevance. "Known" documents are relevant on all of the relevance dimensions except for novelty. Likewise "difficult" documents are relevant on all dimensions except for understandability. "Useless" documents do not fit into the three other categories, e.g., because they are off topic.

Procedure. An experimental run proceeded as follows. First, the eye tracker needed to be calibrated using a 9-point calibration method. Next, the participant read through

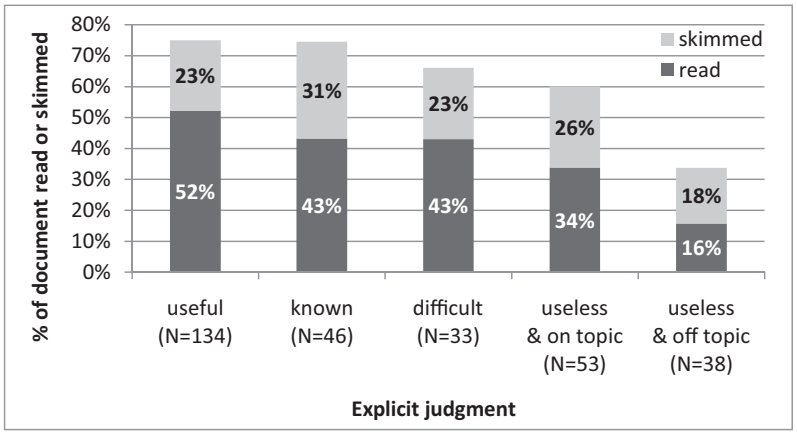

Fig. 3. Percentage of read text for documents broken down by explicit judgments.

the task description informing about the first and the imaginary second phase of the experiment, and also about the explicit categorization scheme. After this introduction, the participant had to look through each of the 16 documents sequentially and provide his or her explicit judgment. A typical experimental run took around 30 minutes.

4.1.2. Apparatus. The experiment was performed on a Tobii[2] 1750 eye tracker at a screen resolution of 1280x1024 pixels. The eye tracker has a tracking frequency of 50Hz and an accuracy of 0.5° of visual angle. For calibrating the device we used the software ClearView from Tobii. To analyze the eye movements, detect reading behavior, and compute the gaze-based measures, we applied our own implemented software using the Tobii SDK. All documents were displayed by the Firefox[3] browser for which we implemented a plug-in asking for explicit judgments every time a document was exited.

4.1.3. Participants. Nineteen participants (10 female) took part in the experiment and produced valid eye tracking data. All of them were undergraduate or graduate students attracted by advertisements for the study, and they had German as their native language. They ranged in age from 22 to 32 years (mean = 24.2, $\sigma = 3.2$) and had a variety of different majors.

4.1.4. Measures. Using our reading and skimming detection method, we analyzed the participants' eye movement patterns while reading the assigned documents. Based on the eye movements belonging to reading behavior (as detected by our method), we computed the previously described gaze-based measures. Therefore, the data for the analysis had the following structure: for every document and participant we had an explicit judgment and a set of gaze traces that were used to compute gaze-based measures.

4.2. Results

4.2.1. Distribution of Reading Behavior. Figure 3 shows the amount of reading and skimming behavior on documents broken down by explicit judgment. The number N of documents in each category is given in the category descriptors.

[2]http://www.tobii.com.
[3]http://www.mozilla.com/firefox/.

Table I.
Means and standard deviations for the measures "average forward saccade length", "thorough reading ratio", and "coherently read text length" split by explicit judgment

relevance judgment	mean forw. sacc. length		thorough reading ratio		coher. read text length	
	mean	σ	mean	σ	mean	σ
relevant	9.36 letters*	2.84	0.67*	0.36	465 letters*	430
known	9.99 letters	3.20	0.56	0.30	409 letters	381
difficult	9.80 letters*	3.03	0.62*	0.39	426 letters*	438
useless & on topic	10.41 letters	3.57	0.54*	0.43	317 letters*	347
useless & off topic	10.95 letters	3.23	0.41	0.42	233 letters	277

It should be noted that the category "useless" is further split into "useless & on topic" and "useless & off topic" where on topic refers to documents that have at least something to do with the main topic of the task (14 of the 16 documents), and where off topic refers to the 2 completely off-topic documents. Topicality was determined by the authors, not by the participants. All participants judged the two off-topic documents and some additional on-topic documents as useless.

The differences in the amount of read plus skimmed text of each document are clearly statistically significant between the category "useless & off topic" and all remaining categories at the 0.01 level (e.g., t-test for "useful" vs. "useless & off topic": $t(170) = 9.6$, $\alpha < 0.01$). However, the differences in the amount of reading plus skimming between the categories "useful" and "known" and between "difficult" and "useless & on topic" are not statistically significant.

The ratio between reading and skimming behavior is different for the 5 categories in Figure 3. Documents being on topic were mostly read instead of skimmed. On the contrary, useless off topic documents lead to slightly more skimming than reading behavior (however, not statistically significant).

Discussion. The combined amount of reading and skimming behavior on the 14 on-topic documents is surprisingly similar. Participants read or skimmed only 15% less of on-topic documents judged useless than of documents judged useful. Also, there does not seem to be a difference in the amount of reading or skimming on useful and known documents. Only the difference to useless off-topic documents is clear and statistically significant.

The findings show that documents being on topic in some way do induce reading behavior of a similar amount. It seems that users keep reading or skimming if documents are on topic and if they can be expected to contain useful information. A reason for this is to a great extend the characteristic of the task. Since users were asked to assess entire documents, they tended not to skip much text. In contrast to topicality, the factors novelty and understandability seem to have only slight influence on the amount of reading behavior.

The ratio between reading and skimming behavior while viewing documents seems to be related to the participants' final explicit judgments. One of our measures, the thorough reading measure, is computed as this ratio and will be analyzed in more detail in the following.

4.2.2. Gaze Measures to Estimate Relevance

Absolute measures. Table I presents means and standard deviations for the measures "average forward saccade length," "thorough reading ratio," and "coherently read text length" across participants and documents. An asterisk signifies statistical significance of the differences to the category "useless & off topic" (student's unpaired, two-sided t-test using a significance level of $\alpha < 0.01$).

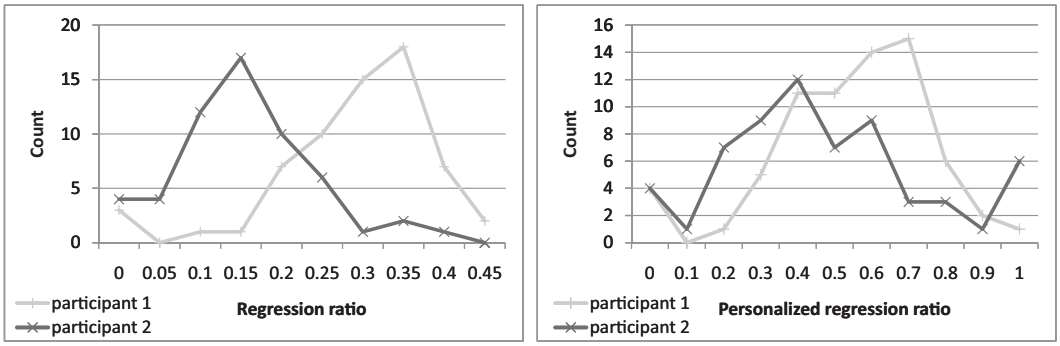

Fig. 4. Absolute and personalized value distribution of "regression ratio" for two participants.

Table II.
Absolute and personalized means and standard deviations for the measures "regression ratio" and "average fixation duration" split by explicit judgment

relevance judgment	regression ratio				average fixation duration			
	absolute		personalized		absolute		personalized	
	mean	σ	mean	σ	mean	σ	mean	σ
relevant	0.16	0.13	0.42*	0.27	219ms	67	9.55*	0.22
known	0.16	0.11	0.44*	0.34	228ms	75	0.52	0.23
difficult	0.16	0.13	0.42*	0.26	202ms	71	0.52	0.22
useless & on topic	0.17	0.16	0.39	0.30	215ms	68	0.50	0.25
useless & off topic	0.14	0.18	0.34	0.33	208ms	81	0.48	0.27

Concerning the two measures "average fixation duration" (overall mean = 216ms, $\sigma = 71$) and "regression ratio" (overall mean = 0.16, $\sigma = 0.14$), we could not determine any significant differences with respect to the five relevance categories.

Personalization. We were surprised not to see any significant differences for fixation duration and regression ratio since previous literature used these measures to estimate relevance [Ohno 2004; Xu et al. 2009; Moe et al. 2007; Brooks et al. 2006]. Therefore, we analyzed the individual distributions of the measured values for each participant.

It turned out that there are big individual differences. Figure 4 (left) shows the absolute value distribution for the measure "regression ratio" for two participants. Participant 1 typically performs more regressions during reading (mean = 29%) than participant 2 (mean = 17%). As we know from the literature [Rayner 1998], these differences do not necessarily mean that one of the participants found the documents generally more relevant than the other. Rather, these types of individual differences are highly dependent on the typical reading strategy and expertise of the reader.

In order to achieve comparability of the measured values across participants, we personalized them by applying the personalization method from Section 3.4. The resulting distributions for "regression ratio" for participants 1 and 2 are presented in Figure 4 (right).

Personalized measures. Table II presents absolute as well as personalized means and standard deviations for the measures "regression ratio" and "average fixation duration," calculated across participants and documents. Again, an asterisk signifies statistical significance of the difference to the category "useless & off topic" (determined by a student's unpaired, two-sided t-test using a significance level of $\alpha < 0.01$).

While none of the differences concerning the category "useless & off topic" are statistically significant with respect to the absolute values for both measures, some of

the differences are significant after personalizing them. Particularly for the measure "regression ratio" a clear trend can be observed. For "average fixation duration," only a slight tendency can be found.

Discussion. The most expressive measure with respect to perceived relevance is based on the length of coherently read text. Participants read about 4 coherent lines of text (corresponding to 465 characters in out setting) in paragraphs that belonged to documents judged useful. In contrast, they only read 2 coherent lines of text (233 characters) in paragraphs belonging to useless off-topic documents. This difference shows how much users tend to skip text if it does not appear to be relevant.

There is evidence that mean forward saccade length and thorough reading ratio are related to relevance of viewed text. On the one hand, the length of left-to-right saccades decreases with perceived relevance of text. On the other hand, the percentage of read (vs. skimmed) text increases with perceived relevance.

With respect to "regression ratio" and "average fixation duration" we find that the measures are not very discriminative when considering values on their absolute scale. There exist great differences across participants with respect to the individual distributions for these measures. But after individually personalizing the measures, we see clear signals for "regression ratio": the percentage of regressions during reading generally increases with perceived relevance. However, concerning "average fixation duration", we only see a very slight trend that duration increases with perceived relevance. But there is little evidence for it.

The results show that mean forward saccade length, thorough reading ratio, and regression ratio are all influenced by perceived relevance of read text. However, the finding that fixation duration is the least expressive measure is somewhat surprising since it is most frequently used in previous research as a simple indicator for relevance [Ohno 2004; Xu et al. 2009].

4.3. Conclusion

The study shows that reading and skimming behavior is strongly influenced by relevance and especially by topicality of text. Whether a document is on or off topic is clearly indicated by the amount of reading behavior. Therefore, the amount of reading behavior alone gives a good first approximation of whether some text has been relevant to the user.

Relevance, and topicality in particular, are also reflected in most of our analyzed eye movement measures. "Coherently read text length," "thorough reading ratio," and "regression ratio" increase with perceived relevance of the read text. "Mean forward saccade length" decreases with perceived relevance. Interestingly, we did not find strong evidence that fixation duration is influenced by relevance.

Even in our rather homogeneous set of participants (i.e., young students at university) we found large individual differences across participants. Some eye movement measures (particularly "regression ratio") had different, individually specific distributions and mean values. Therefore, any signals in these measures with respect to relevance were overwhelmed by the noise that was introduced by individual differences. However, these individual differences seem to be systematic, so that personalizing and normalizing the measures greatly increased their expressivity with respect to relevance.

The found relationships of eye movement measures with perceived relevance of the text may be very useful for generating relevance feedback on the fly for improving and personalizing information retrieval. The next study will further validate the relationships and measure their impact on information retrieval personalization.

5. STUDY 2: READING BEHAVIOR AS IMPLICIT FEEDBACK FOR SEARCH

The first study demonstrates that detecting reading behavior is very valuable and useful, and that there clearly are signals in eye movements with respect to individually perceived relevance. However, the experimental scenario was very specific in that the assigned documents were all either completely on or off topic and that they were all very short and homogeneous. In addition, the given task, that is, assessing the relevance and usefulness of documents, could have had a high impact on the observed reading behavior. In more realistic settings, tasks can be much more directed to finding specific information, and documents can be much longer and can contain some parts that are relevant and other parts that are irrelevant to the task. Also, the first study comprised a document assessment task which presumably leads to just one specific type of reading behavior. It can be expected that reading behavior on such structurally different documents and for a different task is more diverse.

For this second study we had two fundamental goals. First, we wanted to validate the findings from the previous study for different task and document characteristics, and second, we wanted to prove that gaze-based feedback is not only full of potential for estimating relevant document parts, but that it is actually very useful as implicit relevance feedback on the segment level for information retrieval.

In specific, concerning validating study 1 results and estimating relevant document parts, the aims for this study are (1) to study reading behavior in general on long documents that contain relevant as well as irrelevant parts, (2) to find out whether the relations between eye movement measures and relevance found in the previous study are also valid for different document structures and a more goal-directed task, and (3) to determine the influence of an additional factor on reading behavior, that is, familiarity with the document.

For the second fundamental goals, in order to assess the effect of such gaze-based relevance feedback on information retrieval, we compare it to baseline implicit feedback methods in a search personalization scenario comprising query expansion and re-ranking of search results. (Results for this part of the study have partly been published in Buscher et al. [2009b] before.) The baseline methods do not comprise any information from eye tracking at all. They rather create implicit feedback incorporating information from the user query and the contents of the documents the user has viewed immediately before querying the search engine. To quantify the effect of our implicit feedback methods on search result re-ranking, we also compare them with the original ranking from the commercial Web search engine Live Search.[4]

The information retrieval scenario we are using to evaluate such implicit relevance feedback is as follows. We assume that a user has an information need in mind and looks through some documents in order to find relevant information. Next, he or she queries a search engine to find more related information. At this point, we aim at personalizing the result ranking that is produced by the search engine since we know which documents the user has viewed before and how he or she has viewed them. Thus, we use the user's viewing behavior as implicit feedback for the next query and can either expand the query with additional relevant terms or re-rank the result list generated by the search engine. The search engine should then return more relevant results to the user.

5.1. Methods

5.1.1. Experimental Design and Procedure. We designed an eye tracking experiment where participants had to look through four long documents in order to read up on

[4]http://www.live.com/.

a given topic (*the reading phase*), and then search for more related information using a search engine and provide us with explicit relevance ratings for the returned search results (*the searching phase*).

The Tasks. The two topics we gave to each participant were

(1) *perception* (about the variety of perceptual organs of animals and how they work) and
(2) *thermoregulation* (about thermoregulation mechanisms of animals used to control their body temperature).

In order to read up on each topic, the participants had to look through the four provided long documents. They were not asked to provide any explicit relevance judgments of any kind while viewing the documents.

After viewing the four documents with one of the topics in mind they had to pose three different queries to the search engine (searching phase). The first query should always find more related information about the general topic of the task. The two remaining queries should return information about two specific subtopics. For example concerning the topic *perception*, the participants should find more information about *visual* and *magnetic perception* in particular. The general topics for the queries were told orally to the participants. They were free in formulating them.

For each of their queries, the participants were asked to provide explicit relevance judgments for the top 20 of the results on a 6 point relevance scale ranging from "−−−" to "+++". To judge a result, they were told not rely on the search result abstract visible on the search engine results page but to open the document and look through it. In contrast to study 1, this more common relevance rating scale was employed to be able to measure the performance of different implicit feedback methods using standard metrics.

In order to keep them focused on the respective topic of the task, they had moderate time pressure, that is, about 15 minutes for the reading phase and an additional 15 minutes for the searching phase for each task.

The Documents for the Reading Phase. For reading up on the two topics concerning both tasks during the reading phase, the participants had to use the same four assigned documents. Each document was taken from the German version of Wikipedia[5] and was about a different animal species, i.e., about snakes, bees, dogs, and seals. The original Wikipedia articles were modified slightly so that all documents had a comparable length of around 4200 words and so that approximately 6% of the contained text was relevant for each of the two task topics. Each article dealt with a great variety of aspects concerning the respective animal species. However, each article contained some relevant subsections with respect to both topics. We placed these subsections in mostly random positions of the document, with the exception that they never occured in the very beginning or the very end of a document. Since we knew about the positions of the relevant sections, we did not need explicit relevance judgments for text passages from the participants.

The documents were mostly composed of plain text at a font size of 12pt. All of the documents contained some pictures. We removed the table of contents from the original Wikipedia articles and any other means of navigating within the document. Section headings were kept, but not in bold font.

Reranking Procedures. The Web search interface they had to use was modified by us so that we could analyze personalization techniques based on implicit feedback collected from the reading phase.

[5]http://de.wikipedia.org.

To assess the effect of the implicit feedback methods on the final ranking of the search results, we applied two different procedures: *query expansion* and *reranking*. The two procedures were not applied simultaneously but we rather split the participants of the study in two separate groups. The first group of 17 participants used a query interface that comprised reranking. For the second group of 15 participants we applied query expansion. However, the search interfaces looked exactly alike and it was not possible for the participants to identify which group they belonged to.

Task Procedure. To make the experiment look as realistic as possible to the participants, the story and the task protocol were designed as follows. The participants had to imagine being journalists having to write articles for a newspaper. We provided them with a simulated email from their imaginary advisor stating the topics (i.e., perception and thermoregulation) they had to write their next newspaper article on. The emails contained our four preselected documents as attachments which should help them to get started with the topic. After having looked through the four documents within about 15 minutes of time (reading phase), the participants had to employ our search interface to find more related information for another 15 minutes (searching phase).

After having finished with the reading phase and the searching phase for one topic, the participants had to repeat the exact same procedure for the other topic. Using the same four documents for both topics guaranteed that the participants were familiar with the documents' structure for the second reading session. However, the Web search session in between the two reading sessions ensured that the participants could not exactly remember the structure of the documents anymore. Half of the participants started the experiment using the topic *perception* and the other half started with *thermoregulation*.

Completing the task for both topics together usually took one hour per participant. The eye tracker was used during the reading phase only.

In summary, the procedure looked as follows from the participants' viewpoint.

(1) They were informed about the task and the current topic in a simulated email.
(2) They had to look through four preselected long documents that contained some parts relevant to their topic.
(3) They had to query a Web search interface three times to find more related information.
(4) For each query they had to provide relevance ratings for the top 20 returned results.

This procedure was repeated twice, that is, for both topics.

5.1.2. Implicit Feedback Methods. For personalization purposes of search engine result rankings, we analyze and compare five different implicit feedback methods. Overall, we evaluated implicit feedback methods based on all investigated eye movement features. However, since they turned out to have very similar performance, we report only the results from the three most interesting such implicit feedback methods. In all cases, implicit feedback is generated based on the documents the participant viewed during the reading phase, that is, immediately before querying the search engine. All of the implicit feedback methods extract a list of terms most characteristic to describe the user context based on their feedback source. These terms are then either used to expand the user query or to re-rank results from the Web search engine.

Method *Reading*. This method is based on eye tracking and reading detection. It first concatenates all documents viewed by the user resulting in one large context document d. Second, it creates a filtered context document d_P by removing all text parts from

d that have neither been read nor skimmed. Third, most characteristic terms are extracted from document d_P simply based on their $tf \times idf$ scores.[6]

Method *ReadLength(l)*. This method is an extension of the previous method. It is not solely based on the reading and skimming detection method but also incorporates information from the measure "coherently read text length" to decide whether a certain part of read text should be counted as positive or negative relevance feedback. As demonstrated in the previous two studies, length of coherently read text is strongly related to relevance. Therefore, terms occurring in long coherently read text parts should be better descriptors of the user's topic of interest than terms in short read text parts.

In more detail, the method uses the same large context document d as in the previous method. It creates two filtered context documents d_P and d_N. Document d_P contains all parts of d that have been read or skimmed and that have a "coherently read text length" of at least l characters. On the contrary, d_N contains all parts of d that have been read or skimmed but whose "coherently read text length" has been smaller than l characters. Next, scores for all terms w in d_P are computed using the following formula.[6]

$$\frac{tf(w, d_P)}{tf(w, d_P) + tf(w, d_N)} \times idf(w). \tag{1}$$

$tf(w, d_X)$ denotes the term frequency of term w in document d_X.

Based on this computation, terms that are very specific for long read text parts will get higher scores than terms that also appear in short read text parts.

Method *ReadExtremes(l_1, l_2)*. This method is an immediate extension of the previous method. It also uses the same large context document d and creates two filtered context documents d_P and d_N. However, document d_P contains all parts of d that have been read or skimmed and that have a "coherently read text length" of at least l_2 characters, and d_N contains those that are smaller than l_1 characters. Scores for all single terms included in d_P and d_N are computed as in the previous method.

The difference to the method *ReadLength(l)* is that only these read text passages are used as positive or negative feedback that are close to both extremes of the distribution in terms of coherently read text length. Read text passages with an intermediate length are ignored.

Method *FullDocument*. As a simple baseline, this method just uses information about what documents have been viewed before issuing a query. Therefore, it just extracts the most characteristic terms of the four documents we provided to the participants. The documents are not filtered any further. Most characteristic terms are determined according to their $tf \times idf$ values.[6]

Method *QueryFocus*. This method can be seen as a pseudorelevance feedback method on the segment level of documents. It takes into account which documents the participant has viewed before issuing the query, and it also considers the issued query itself. The query is used to find these parts of the previously viewed documents that are relevant to the query at hand.

In more detail, the context document d (as before) is split into a set of paragraphs. For each paragraph, a ranking score is determined with respect to the user query using Lucene[7] and a BM25 ranking function. Then, the best paragraphs are selected, that is, the paragraphs achieving a ranking score not worse than half of the ranking

[6]For the computation of idf values Wikipedia was used as a document corpus.
[7]http://lucene.apache.org/.

score achieved by the best matching paragraph. A virtual document d_P is created concatenating all selected paragraphs. As before, the most characteristic terms are then extracted from d_P based on their $tf \times idf$ scores.

5.1.3. Personalization Procedures. There were mainly two reasons why we decided to examine the effect of the implicit feedback methods using two procedures, i.e., reranking and query expansion.

First, reranking has the technical advantage that implicit feedback methods (i.e., the method *ReadLength(l)* with different parameter settings l can be examined even after an experimental run is over. All the data needed to compute the quality of feedback methods can be stored offline: eye tracking data for generating feedback, and relevance judgments for the top k results of the original user query. This is not possible using the query expansion procedure since every new feedback method results in a new unique query. Every unique query can result in a completely different set of result documents for which we needed to ask the participant again to provide relevance judgments.

Second, query expansion has the potential advantage of higher personalization impact. Since every expanded query is unique, it can retrieve new documents from the Web and can exclude others compared to the nonexpanded user query. Therefore, we assumed that expanding the original user query with additional terms would have much larger effects on the quality of the final ranking than just reranking the top k results for the original user query.

Reranking Procedure. For this procedure, each user query was first submitted to the Live Search engine using their SDK in order to retrieve the top 20 result Web pages. The pages were automatically downloaded and stored locally. The results were presented to the participants in randomized order (whom we informed about the randomization). After collecting relevance judgments for all 20 results from the participant, we had all necessary data for offline posteriori reranking:

—the original user query,
—the top 20 resulting Web documents,
—relevance judgments for these documents,
—eye tracking data specifying how the user has viewed the four assigned documents during the reading phase of an experimental session.

To re-rank the top 20 results we performed several steps: First, Lucene[8] was used to generate an index over the top 20 result documents. Second, each implicit feedback method to evaluate was executed on the four provided, previously read documents resulting in a list of characteristic terms paired with their achieved scores. Third, the original user query was expanded by the extracted terms weighted by their scores. Weights were also added to the original user query terms so that the weight of all expansion terms together accounted for 60% of all term weights. Overall, an expanded query contained at most 19 terms (due to technical reasons). Fourth, Lucene was applied to rerank the top 20 original result documents using the expanded query.

Query Expansion Procedure. With this procedure we aimed at comparing some of the feedback methods from above. Based on the results we got from the reranking procedure (which was analyzed first; see the result section for further details), we decided to consider the methods *Reading* as a representative gaze-based feedback method and *QueryFocus* as the best non-gaze-based method. We further wanted to compare the results to the original ranking from the Web search engine Live Search for the original user query.

[8]http://lucene.apache.org/.

To expand and evaluate the queries, we proceeded as follows. First, each implicit feedback method to evaluate extracted a list of most characteristic terms from the four assigned documents. Second, each user query was expanded with the top m extracted terms so that the resulting expanded query had the following syntax:

$term_{U1} \ldots term_{Un}$ ($term_{E1}$ OR ... OR $term_{Em}$)

Hence, the original user search terms $term_{Ui}$ appeared in a space-separated list whereas the extracted terms $term_{Ei}$ were connected with OR operators. Due to technical reasons of the Web search engine Live Search, a query contained at most 19 terms again, that is, $n+m \leq 19$. Third, each expanded query was submitted to Live Search. Fourth, the top returned results from the original user query and all expanded query variants were merged together and presented to the participant in randomized order. All presented documents were then judged by the participant so that we could compute and compare the quality of the returned result lists from the original and the expanded queries.

5.1.4. Apparatus. The experiment was performed using the same Tobii 1750 eye tracker that was applied before in the first study (Section 4.1.2). The screen resolution was kept at 1280x1024 pixels. Before starting an experimental run, the device was calibrated using ClearView and a 9-point calibration method. To detect reading, we used the same implemented methods and techniques as before. The documents and the search interface were presented in a Firefox browser.

5.1.5. Participants. The study was performed by 32 participants (6 female). Most of them were undergraduate students, some were graduates taking a variety of different majors. All of them had German as their native language. Their age ranged between 20 and 28 years (mean = 22.7, $\sigma = 1.7$).

A first group of 17 participants was assigned to the re-ranking procedure. The remaining 15 participants were assigned to the query expansion procedure.

5.1.6. Measures. Reading-related measures. As in the first study, we applied our reading and skimming detection method to classify gaze traces as either belonging to reading, to skimming, or to non-reading-related behavior. Based on the eye movements belonging to reading or skimming, we computed the same gaze-based measures as before on the level of paragraphs in the documents. Hence, the structure of the data used for the analysis was as follows: for every participant and every paragraph in each document we had a set of gaze traces and corresponding eye movement measures and we knew whether the respective paragraph was relevant to the topic of the task at hand.

Information retrieval measures. To measure the quality of the different search result lists generated based on the different implicit feedback methods, we computed three well accepted information retrieval metrics, each focusing on different aspects:

—*Precision at K*. $P(K)$ computes the fraction of relevant documents within the top K results. However, the order of the top K documents does not matter. Since this measure needs binary relevance judgments, we split the 6-point relevance scale that was used in the experiment into two groups – positive and negative.

—*DCG at K*. The Discounted Cumulative Gain [Järvelin and Kekäläinen 2000] $DCG(K)$ is different from $P(K)$ in that it considers the order of the top K results and in that it is not bound to binary relevance judgments. It is computed as

$$DCG(K) = \sum_{j=1}^{K} \left(2^{r(j)} - 1\right) / log(1+j) \tag{2}$$

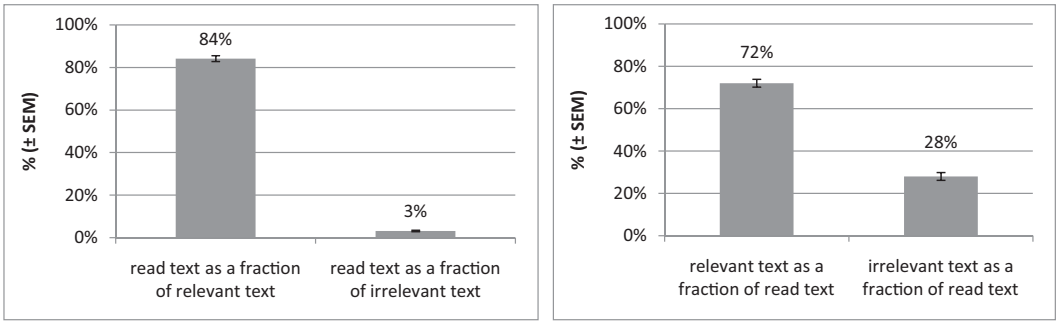

Fig. 5. Distribution of reading behavior. Left: read text as a fraction of relevant/irrelevant text. Right: relevant/irrelevant text as a fraction of read text.

Table III. Means and Standard Deviations for All Eye Movement Measures from the Second Study

measure	relevanti		rrelevant	
	mean	σ	mean	σ
mean forward saccade length *	7.8 letters	3.62	11.8 letters	7.68
thorough reading ratio *	0.53	0.35	0.40	0.45
coherently read text length *	351 letters	278	121 letters	163
regression ratio (personalized) *	0.41	0.24	0.30	0.33
average fixation duration (personalized)	0.51	0.22	0.49	0.25

$r(j)$ corresponds to the relevance score of the j^{th} result entry. In our case, the judgments $+++$, $++$, $+$ led to relevance scores of 3, 2, and 1, respectively, whereas all negative judgments got a relevance score of 0.

—*MAP*. Mean Average Precision is a single value metric to measure the quality of a set of result lists. It is calculated as the mean of average precision values for all result lists in the set. Concerning one result list, average precision is computed as the mean of the $P(K)$ values after each relevant document was retrieved.

5.2. Results: Reading Behavior

5.2.1. Validation: Distribution of Reading Behavior. Figure 5 shows the distribution of reading behavior on the four assigned documents with respect to relevance of the read text parts. In general, 84% of the text that was relevant has been read (Figure 5 left) and 72% of all text that has been read was relevant (Figure 5 right).

Discussion. The results shows that reading behavior is very focused on the relevant parts of the text. Irrelevant parts are largely ignored. Compared to the previous study, the tendency to skip irrelevant parts is much greater (i.e., 34% of irrelevant off-topic documents have been read or skimmed in the previous study, while only 3% or irrelevant parts have been read or skimmed in this study). One of the reasons for this tendency is the different document structure in this study: most of the document parts were irrelevant to the topic of the task. In contrast, most of the documents in the previous study were completely on topic and most of them were judged relevant.

Almost 3/4 of all text that has been read or skimmed was relevant. This shows that reading behavior alone is already a good indicator for relevance.

5.2.2. Gaze-Based Measures. Table III gives an overview of mean values and standard deviations for the measures that were also discussed in the first study. "Regression ratio" and "average fixation duration" are both computed in their individually

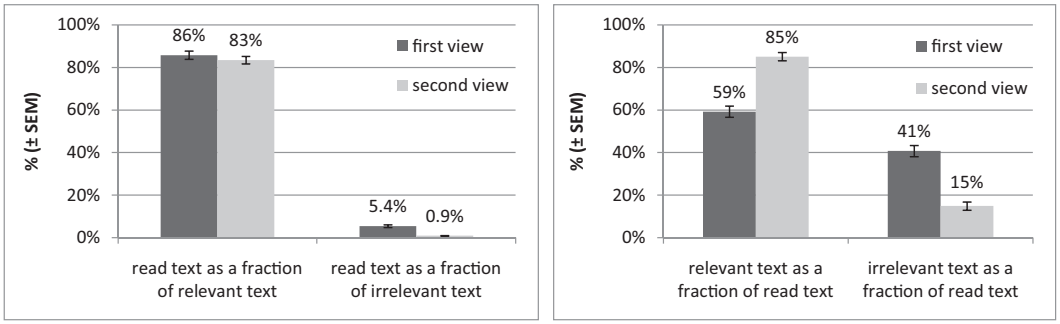

Fig. 6. Distribution of reading behavior split by first and second view of each document. Left: read text as a fraction of relevant/irrelevant text. Right: relevant/irrelevant text as a fraction of read text.

personalized forms, as before. An asterisk signifies that the differences of the means on relevant and on irrelevant text parts are statistically significant (student's unpaired, two-sided t-test using a significance level of $\alpha < 0.01$).

Compared to the results from the first study, we see the exact same trends in this study: The length of saccades is typically smaller on relevant text sections. "Thorough reading ratio," "coherently read text length," and "regression ratio" are increased on relevant parts. Like in the previous study, we could not determine any statistically significant difference in "average fixation duration" with respect to relevance.

However, when comparing the specific mean values of the measures for relevant and irrelevant parts between the two studies, then some differences are evident. "Mean forward saccade length" has a much wider value range in this study. The values for "thorough reading ratio" are more spread in the previous study. "Coherently read text length" is generally shorter in this study. In contrast, the value distribution of personalized "regression ratio" has very similar characteristics in both studies.

Discussion. In general, the results verify and confirm the findings from the first study. They show that reliable signals can be found in the four measures "mean forward saccade length," "thorough reading ratio," "coherently read text length," and "regression ratio" with respect to relevance. Again, "coherently read text length" is the most expressive measure.

However, the mean values for most of the measures seem to depend on further factors related to task and document structure and are different in both studies. Therefore, it might be difficult to find generic classifiers for predicting relevance that work in a variety of different settings since most of the measures are not robust enough and are not independent from task and document structure. In this respect, personalized "regression ratio" seems to be the most robust measure considering both study settings.

5.2.3. Differences Between First-Time and Second-Time Reading. Every document has been presented twice to every participant, once for each of the two topics. Figure 6 shows the distribution of reading behavior again, but now split by first and second view of each document. While a similar amount of the relevant document parts were read during both views, 6 times less irrelevant text was read during the second view (Figure 6, left). This is also reflected in the amount of relevant or irrelevant text as a fraction of all read text (Figure 6, right).

We also analyzed whether there were differences concerning the value distribution of the measures with respect to first and second view of a document. We found a significant difference concerning coherently read text length being longer for first views than for second views, both for relevant and irrelevant text parts (to determine significance we

Table IV. MAP and DCG at $K = 10$ for Different Parameter Settings l for the Methods $ReadLength(l)$ and $ReadExtremes(l_1, l_2)$

Variant	MAP	DCG
ReadLength(0 chars)	0.747	9.14
ReadLength(50 chars)	0.749	9.18
ReadLength(100 chars)	0.733	9.17
ReadLength(150 chars)	0.736	9.15
ReadLength(200 chars)	0.734	9.06
ReadLength(250 chars)	0.740	9.26
ReadLength(300 chars)	0.743	9.18
ReadLength(350 chars)	0.737	9.14

Variant	MAP	DCG
ReadExtremes(100,300)	0.752	9.23
ReadExtremes(100,350)	0.748	9.26
ReadExtremes(100,400)	0.756	9.28
ReadExtremes(100,450)	0.753	9.31
ReadExtremes(150,400)	0.755	9.27
ReadExtremes(50,400)	0.753	9.27

used Bonferroni correction of $\alpha = 0.001$). Beyond that, we could not find any further significant differences between first and second view concerning the remaining measures.

Discussion. The results demonstrate that reading behavior becomes much more focused on relevant document parts if users have already seen a document and are familiar with its structure. This means that reading behavior alone is already very precise in pointing out relevant parts of known documents (with a precision of 85% and a recall of 83%).

Interestingly, most of the eye movement measures are very robust and do not show significant differences with respect to document familiarity. This is important to know when trying to build generic classifiers to determine relevance based on eye movements.

5.3. Results: Implicit Feedback

All 32 participants together issued 192 queries and gave relevance judgments for 3497 result entries. We describe our findings in this section as follows: First, we evaluate the best parameter settings for the implicit feedback methods $ReadLength(l)$ and $ReadExtremes(l_1, l_2)$. Second, we compare the quality of the result lists produced by the different implicit feedback methods, both using the re-ranking as well as the query expansion procedure.

5.3.1. Evaluating Parameter Settings. Based on the data from the re-ranking procedure, we could test several values for the parameter l of the method $ReadLength(l)$ and the parameters l_1 and l_2 of the method $ReadExtremes(l_1, l_2)$. Table IV shows MAP and DCG scores at $K = 10$ after re-ranking the top 20 search results from the Web search engine based on the two implicit feedback method. The best rankings could be achieved with a parameter l = 50 characters, and $l_1 = 100$, $l_2 = 400$ characters.

Discussion. Given the results from the previous analyses, it is not surprising that we can see slight improvements of the ranking quality when filtering read text by coherently read text length. Regarding the method $ReadExtremes(l_1, l_2)$ we find that the best parameters are close to the mean values for coherently text lengths on relevant and irrelevant texts, and overall this method seems to be superior to $ReadLength(l)$. This shows that read text passages of lengths between about 100 and 400 characters introduce a considerable amount of noise and should not be used as positive or negative relevace feedback.

5.3.2. Comparison of Method Performance. The diagrams in Figure 7 show the performances of the different implicit feedback methods with respect to the reranking procedure as well as the quality of the original ranking from the Web search engine. In addition, Table V presents MAP and DCG scores at $K = 10$ both for the reranking and the query expansion procedure. Asterisks denote statistical significance of the

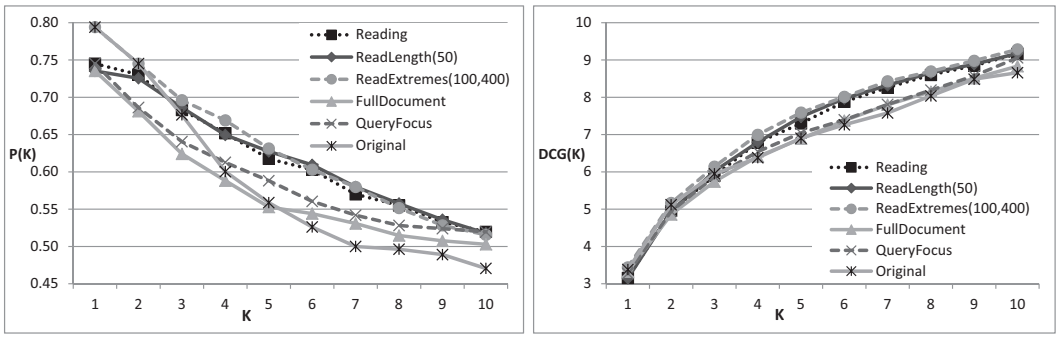

Fig. 7. Precision and DCG at recall levels K for the re-ranking procedure.

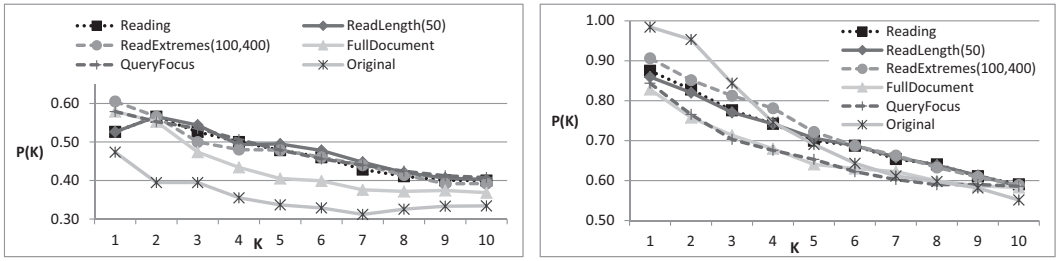

Fig. 8. Precision at different recall levels K for poorly performing queries (left) and highly performing queries (right).

Table V. MAP and DCG at $K = 10$ for the Reranking and Query Expansion Procedure

Variant	Re-Ranking MAP	Re-Ranking DCG	Query Expansion MAP	Query Expansion DCG
(1) Reading	0.748 *4,5	9.17 *4,5		
(2) ReadLength(50)	0.749 *4,5	9.18 *4,5	0.762	8.58 *6
(3) ReadExtremes(100,400)	0.756 *4,5	9.28 *4,5		
(4) FullDocument	0.697	8.83		
(5) QueryFocus	0.709	9.06	0.751	8.21
(6) Original	0.729	8.66	0.751	8.05

differences to the respective numbered methods (determined by a student's paired, two-sided t-test using a significance level of $\alpha < 0.05$).

While the gaze-based feedback methods produced significantly better rankings than the non-gaze-based feedback methods, none of the four implicit feedback methods led to significantly better or worse rankings compared to the original ranking from the Web search engine. Hence, inspired by Agichtein et al. [2006] we further explored the effect of implicit feedback by dividing all user queries into two groups: *poorly performing queries* where the Web search engine produced rankings with MAP scores ≤ 0.7, and *highly performing queries* (MAP > 0.7). The former group contained 38, the latter group 64 queries.

Figure 8 shows precision-recall diagrams for poorly and highly performing queries with respect to the reranking procedure. Table VI provides respective MAP and DCG values at $K = 10$. Asterisks denote statistical significance of the differences as before.

Table VI. MAP and DCG at $K = 10$ for Queries with Poor and High Original Performance

	MAP of original ranking			
	<= 0.7		> 0.7	
Variant	MAP	DCG	MAP	DCG
(1) Reading	0.63 *6	6.64 *4,6	0.82 *4,5	10.67
(2) ReadLength(50)	0.63 *4,6	6.63 *4,6	0.82 *4,5	10.69
(3) ReadExtremes(100,400)	0.63 *4,6	6.49 *6	0.83 *4,5	10.95 *5
(4) FullDocument	0.58 *6	6.00	0.76	10.52
(5) QueryFocus	0.62 *6	6.80 *1,2,4,6	0.76	10.40
(6) Original	0.50	5.18	0.86 *4,5	10.73

Discussion. Largest gains were achieved by method *ReadExtremes*(100, 400), that is, a 8.5% gain in MAP compared to the simple non-gaze-based feedback method *FullDocument*, and a 7.2% gain in DCG compared to the original ranking from Live Search. All three gaze-based feedback methods lead to overall improvements compared to the original ranking from the Web search engine. In contrast, the two non-gaze-based methods could also lead to impairments, especially in the re-ranking scenario.

This difference in quality between the three gaze-based and the two non-gaze-based methods becomes particularly evident with regard to their performance on highly performing queries. The methods *FullDocument* and *QueryFocus* significantly worsened retrieval performance. In contrast, we could not find significant impairments caused by the gaze-based methods. However, with respect to poorly performing queries, all implicit feedback methods greatly improved the quality of the ranking, that is, up to 27% in MAP.

5.4. Conclusion

Overall, the results of this study verify and confirm the findings from the first study and demonstrate the value and usefulness of gaze-based feedback as implicit relevance feedback in information retrieval.

Concerning reading-related measures we found the same relationships as in the previous study, even in a different setting with long documents containing relevant and irrelevant parts, and for a more goal-directed task: First, the amount of reading behavior is strongly influenced by and very focused on relevant text. Second, the analyzed gaze-based measures show the same trends concerning increases or decreases on relevant or irrelevant text. However, the distributions of their absolute values are different from the previous study (i.e., their means). They seem to depend on the general experimental setting including task and document structure. Third, "fixation duration" seems to be indifferent to relevance again; there were again no significant differences to detect in this respect.

The factor of document familiarity has considerable effects on how users view documents. If users have previous knowledge about the document structure, then they are much more focused while reading. But interestingly, document familiarity had no noticeable effects on eye movement measures with respect to relevance. It did not influence the general trends concerning increases or decreases of the measures on relevant text.

An important insight that can be drawn from both studies is that classifiers to detect relevance based on eye movement measures will not be trivial to build and need to consider relative differences within their value distributions. Classifiers that are just based on absolute values of the measures are unlikely to work for two reasons. First, there are great individual differences. They can be accounted for when personalizing

the measures. Second, there are differences with respect to the general setting, that is, task and document structures. How these setting-induced differences can be accounted for is not yet clear; further research is needed here.

Applying gaze-based feedback as relevance feedback in information retrieval turns out to be very useful. Compared to the non-gaze-based feedback methods, it leads to considerable improvements of the search result list quality both using a reranking and a query expansion procedure. Interestingly, the gaze-based methods are even significantly better than *QueryFocus*, which is the only method incorporating information about the current user query and therefore may use document parts directly related to the query as feedback. However, document parts that have been read or skimmed before are evidently more useful as implicit feedback.

Compared to the original result list quality from the Web search engine, the effects of the feedback methods are less distinct. With respect to queries that lead to poor result list qualities from the Web search engine, all implicit feedback methods lead to great improvements. Particularly *QueryFocus* performs at a comparable level as the gaze-based methods. However, when focusing on queries already yielding good quality result lists then both non-gaze-based methods entail considerable impairments. In contrast, the gaze-based methods do not worsen the result list quality much.

Surprisingly, we could not determine great differences between the three gaze-based methods. We expected to see much larger improvements from the method $ReadLength(l)$ and $ReadExtremes(l_1, l_2)$ since they additionally apply the advanced measure "coherently read text length" which has been proven to be most expressive with respect to relevance. This leads to the conclusion that a bit more irrelevant text as context (used by the method *Reading*) does not hurt retrieval performance much. Hence, reading and skimming behavior alone without considering any more advanced gaze-based measures is already a very effective source for implicit relevance feedback for information retrieval.

However, it has to be kept in mind that this is an exploratory study with specific assumptions on task and document characteristics. First, the documents the participants could read had a very similar structure in that their subsections dealt with largely distinct subtopics (as it is typically the case with Wikipedia articles). Relevant and irrelevant text sections may be less differentiable in other types of documents so that the identification of relevant read text sections may be more difficult, and thus, the relevance measures would be more noisy. Second, whereas the goal-directed task of reading up on a topic for information gathering and then searching for more information on the Web is fairly frequent and generalizeable, time pressure can differ vastly. This may have a significant effect on reading behavior that remains to be estimated in future work. Third, as we demonstrated in this study, reading behavior changes considerably with document familiarity. While our study showed that expressive relevance-related eye movement measures are not much affected by changes in short-term document familiarity, there may be considerable effects coming from long-term document familiarity, that is, if the user is familiar with the document for a long time through multiple re-readings.

Yet overall, this study demonstrates the usefulness of gaze-based feedback for information retrieval personalization. There are many areas of application for such rather simple implicit feedback, for instance, as a search filter in order to refind previously read document parts, as a document interaction measure that can be used to highlight document parts the user has paid most attention to, as a source for creating topical contexts for a user that can be applied as implicit feedback for information retrieval, etc. In the following, we give an outlook of a potential future set of applications: attentive documents.

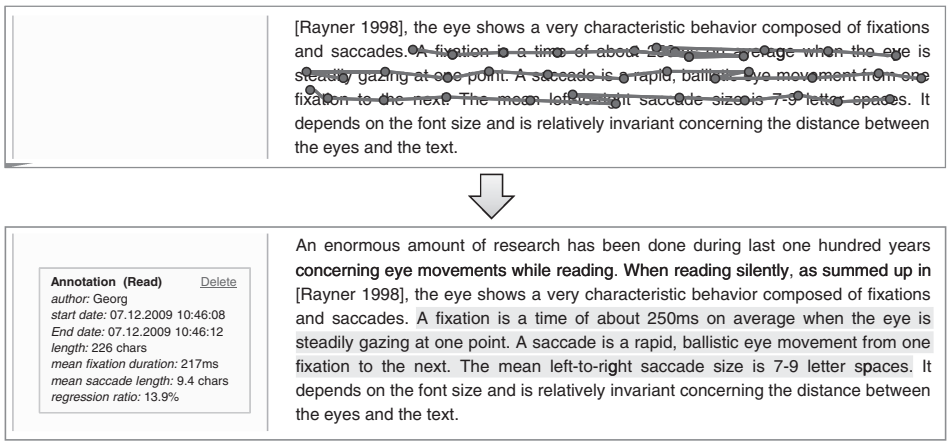

Fig. 9. An attentive Wiki document records what the user is reading (top) and automatically creates annotations containing several eye movement measures (bottom).

6. OUTLOOK: ATTENTIVE DOCUMENTS

Attentive documents are documents that keep track of how they are used. This is not only restricted to editing processes which can already be tracked by modern commercial word processing systems. Since users generally read more than they write, information about how documents and their parts are consumed (and read in particular) can be even more important. Therefore, information about what document parts have been read and how they have been read should be stored in association with the documents. As we demonstrated, eye tracking is overall very well suited to detect reading behavior and to determine differences in reading behavior, e.g., with respect to user-perceived relevance.

We built a prototype system described in detail in van Elst et al. [2008] and Kiesel et al. [2008] illustrating how attentive documents could work. The prototype system itself is a Wiki that allows for semantic annotations of text parts in documents. It can be used to manage the personal information space on the user's desktop. When a user is reading some part of a Wiki document, our reading detection method detects reading or skimming behavior on the fly, computes appropriate gaze-based measures, and then sends this information to the Wiki. Subsequently, the Wiki creates an annotation for the read text passage containing the values from the gaze-based measures. Figure 9 shows an example of such an annotation. Of course, these annotations can be displayed to the user, but they are not visible per se.

Such attentive documents make it possible to design new and innovative applications. For example in the search domain, it is conceivable to use feedback from reading directly as implicit relevance feedback. From this, Web search can effectively be personalized (as demonstrated in study 2). Furthermore, when trying to refind information on the desktop, search filters could be employed that focus and search only in these document parts that have been viewed before by the user. Also, since a system managing attentive documents (such as the local Wiki) knows what the user has read or skipped, it could point out information that is relevant and new to the user, that is, information that has not been read before.

Apart from search, new kinds of applications can be imagined when focusing on the reading process of one particular document. For instance, when a user opens a document that he or she has read a while ago, the system could highlight the parts of

the document that have been most relevant before. This could help the user to refresh his or her memory. Because contextual clues help to find and reconstruct the thoughts one had before, such highlighting of formerly read text parts can speed up the process of recontextualizing. Furthermore, to make reading itself more efficient, the system could notice when the user is just scanning a document very quickly in order to get a rough idea of its contents. In this case, words in the text that do not convey much content (e.g., stop words) can be grayed out so that the scanning process gets more focused and effective [Biedert et al. 2010b]. Additionally, it is also conceivable that the system can detect difficulties of understanding during reading, e.g., on words in a foreign language. Then, it could automatically provide translations or further explanations [Hyrskykari et al. 2003]. Applications aiming at reading entertainment are also imaginable [Biedert et al. 2010a].

The sketched use-cases are just some compelling applications for gaze-based feedback from reading documents. An almost arbitrary number of further applications can be imagined.

7. SUMMARY AND CONCLUSION

In conclusion, it can be stated that gaze-based feedback about what has been read and how it has been read is very valuable to determine whether viewed document parts have been relevant to an individual user. Furthermore, when using this information as implicit relevance feedback it can greatly improve and personalize information retrieval methods.

Reading behavior is very focused on relevant parts of documents, especially when users are working with long documents, and there is strong evidence that individually perceived relevance of read text influences eye movement measures like the number of regressions during reading, the typical length of saccades, etc. We determined the relations between relevance and gaze-based measures and validated them in two studies using different task types and document structures. Interestingly, the popular measure "fixation duration" does not seem to be related to perceived relevance.

We found good relations between gaze-based measures and user-perceived relevance, as well as a great variation in the measures caused by individual differences and differences in task and document structure. Individually personalizing the measures can greatly improve their expressivity and can account for individual differences. However, how variation caused by task and document structure can be accounted for is not clear yet and stays for further research.

Since users typically read in a very focused way, information about what document parts have been read can be used as an indicator of relevance. We demonstrated its usefulness as implicit relevance feedback for personalizing Web search in a further study. The results of the study show that gaze-based feedback is much more effective than non-gaze-based relevance feedback baselines.

Finally, in an outlook, we sketched potential innovative applications based on attentive documents, i.e., documents that keep track of how they have been read by the user.

ACKNOWLEDGMENTS

We thank Tristan King for his help improving the style of the paper.

REFERENCES

AGICHTEIN, E., BRILL, E., AND DUMAIS, S. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, New York, NY, 19–26.

AHN, J. W., BRUSILOVSKY, P., HE, D., GRADY, J., AND LI, Q. 2008. Personalized web exploration with task models. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. ACM, New York, NY, 1–10.

AJANKI, A., HARDOON, D., KASKI, S., PUOLAMÄKI, K., AND SHAWE-TAYLOR, J. 2009. Can eyes reveal interest? Implicit queries from gaze patterns. *User Model. User-Adapt. Interact. 19*, 307–339.

BALATSOUKAS, P. AND RUTHVEN, I. 2010. The use of relevance criteria during predictive judgment: An eye tracking approach. *Proc. Amer. Soc. Info. Sci. Techn. 47,* 1, 1–10.

BIEDERT, R., BUSCHER, G., AND DENGEL, A. 2010a. The eyebook – using eye tracking to enhance the reading experience. *Informatik-Spektrum 33,* 3, 272–281.

BIEDERT, R., BUSCHER, G., SCHWARZ, S., HEES, J., AND DENGEL, A. 2010b. Text 2.0. In *CHI'10: Extended Abstracts on Human Factors in Computing Systems*. ACM Press, New York, NY, 4003–4008.

BROOKS, P., PHANG, K. Y., BRADLEY, R., OARD, D., WHITE, R., AND GUIMBRETIERE, F. 2006. Measuring the utility of gaze detection for task modeling: A preliminary study. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'06)*. (Workshop on Intelligent User Interfaces for Intelligence Analysis).

BUSCHER, G., CUTRELL, E., AND MORRIS, M. R. 2009a. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY, 21–30.

BUSCHER, G., DENGEL, A., AND VAN ELST, L. 2008a. Eye movements as implicit relevance feedback. In *CHI'08: Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, 2991–2996.

BUSCHER, G., DENGEL, A., AND VAN ELST, L. 2008b. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. ACM, New York, NY, 387–394.

BUSCHER, G., VAN ELST, L., AND DENGEL, A. 2009b. Segment-level display time as implicit feedback: a comparison to eye tracking. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*. ACM, New York, NY, 67–74.

CHEN, Z. AND XU, Y. 2005. User-oriented relevance judgment: A conceptual model. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*. IEEE Computer Society, Los Alamitos, CA, 101.2.

CLAYPOOL, M., LE, P., WASED, M., AND BROWN, D. 2001. Implicit interest indicators. In *Proceedings of the 6th International Conference on Intelligent User Interfaces (IUI'01)*. ACM Press, New York, NY, 33–40.

COLE, M. J., GWIZDKA, J., BIERIG, R., BELKIN, N. J., LIU, J., LIU, C., AND ZHANG, X. 2010. Linking search tasks with low-level eye movement patterns. In *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics (ECCE'10)*. ACM, New York, NY, 109–116.

CUTRELL, E. AND GUAN, Z. 2007. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*. ACM Press, New York, NY, 407–416.

DAVENPORT, T. H. AND BECK, J. C. 2001. *The Attention Economy: Understanding the New Currency of Business*. Harvard Business School Press.

VAN ELST, L., KIESEL, M., SCHWARZ, S., BUSCHER, G., AND LAUER, A. 2008. Contextualized Knowledge Acquisition in a Personal Semantic Wiki. In *Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW'08)*. Springer, Lecture Notes in Computer Science vol. 5268, 172–187.

FOX, S., KARNAWAT, K., MYDLAND, M., DUMAIS, S., AND WHITE, T. 2005. Evaluating implicit measures to improve web search. *ACM Trans. Inform. Syst. 23,* 2, 147–168.

GOLOVCHINSKY, G., PRICE, M. N., AND SCHILIT, B. N. 1999. From reading to retrieval: freeform ink annotations as queries. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*. ACM Press, New York, NY, 19–25.

GYLLSTROM, K. 2009. Passages through time: chronicling users' information interaction history by recording when and what they read. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI'09)*. ACM, New York, NY, 147–156.

HILL, W. C., HOLLAN, J. D., WROBLEWSKI, D., AND MCCANDLESS, T. 1992. Edit wear and read wear. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'92)*. ACM Press, New York, NY, 3–9.

HYRSKYKARI, A., MAJARANTA, P., AND RÄIHÄ, K.-J. 2003. Proactive response to eye movements. In *Proceedings of the International Conferece on Human-Computer Interaction (INTERACT'03)*. 129–136.

JÄRVELIN, K. AND KEKÄLÄINEN, J. 2000. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00)*. ACM, New York, NY, 41–48.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inform. Syst. 25,* 2.

Kelly, D. and Belkin, N. J. 2001. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*. ACM, New York, NY, 408–409.

Kelly, D. and Belkin, N. J. 2004. Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*. ACM Press, New York, NY, 377–384.

Kelly, D. and Teevan, J. 2003. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum 37,* 2, 18–28.

Kiesel, M., Schwarz, S., van Elst, L., and Buscher, G. 2008. Using attention and context information for annotations in a semantic wiki. In *Proceedings of the 3rd Semantic Wiki Workshop (SemWiki'08)*.

Liversedge, S. P. and Findlay, J. M. 2000. Saccadic eye movements and cognition. *Trends Cogn. Sci. 4,* 1, 6–14.

Loboda, T. D., Brusilovsky, P., and Brunstein, J. 2011. Inferring word relevance from eye-movements of readers. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI'11)*. ACM, New York, NY, 175–184.

Majaranta, P. and Räihä, K.-J. 2002. Twenty years of eye typing: systems and design issues. In *Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA'02)*. ACM, New York, NY, 15–22.

Melucci, M. and White, R. W. 2007. Discovering hidden contextual factors for implicit feedback. In *Proceedings of the CIR'07 Workshop on Context-Based Information Retrieval* (in conjunction with CONTEXT'07).

Moe, K. K., Jensen, J. M., and Larsen, B. 2007. A qualitative look at eye-tracking for implicit relevance feedback. In *Proceedings of the 2nd International Workshop on Context-Based Information Retrieval*. B.-L. Doan, J. Jose, and M. Melucci, Eds., 36–47.

Morita, M. and Shinoda, Y. 1994. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*. Springer, 272–281.

Ohno, T. 2004. Eyeprint: support of document browsing with eye gaze trace. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI'04)*. ACM, New York, NY, 16–23.

Puolamäki, K., Salojärvi, J., Savia, E., Simola, J., and Kaski, S. 2005. Combining eye movements and collaborative filtering for proactive information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*. ACM Press, New York, NY, 146–153.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psych. Bull. 124,* 3, 372–422.

Rocchio, J. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall, 313–323.

Salton, G. and Buckley, C. 1990. Improving retrieval performance by relevance feedback. *J. Amer. Soc. Inf. Sci. 41,* 4, 288–297.

Simon, H. A. 1969. *The Sciences of the Artificial*. MIT Press.

Simon, H. A. 1971. Designing organizations for an information rich world. In *Computers, Communications and the Public Interest*. Johns Hopkins Press, 38–51.

White, R. W. and Kelly, D. 2006. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*. ACM, New York, NY, 297–306.

Wilcox, R. 2005. *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Ed. Elsevier Academic Press.

Xu, S., Jiang, H., and Lau, F. C. 2009. User-oriented document summarization through vision-based eye-tracking. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI'09)*. ACM, New York, NY, 7–16.

Received December 2010; revised June 2011; accepted August 2011