Attention-based Information Retrieval Using Eye Tracker Data

Tristan Miller and Stefan Agne

German Research Center for Artificial Intelligence (DFKI GmbH)
Postfach 20 80, 67608 Kaiserslautern, Germany
Tristan.Miller@dfki.de, Stefan.Agne@dfki.de

Abstract. We describe eFISK, an automated keyword extraction system which unobtrusively measures the user's attention in order to isolate and identify those areas of a written document the reader finds of greatest interest. Attention is measured by use of eye-tracking hardware consisting of a desk-mounted infrared camera which records various data about the user's eye. The keywords thus identified are subsequently used in the back end of an information retrieval system to help the user find other documents which contain information of interest to him. Unlike traditional IR techniques which compare documents simply on the basis of common terms withal, our system also accounts for the weights users implicitly attach to certain words or sections of the source document. We describe a task-based user study which compares the utility of standard relevance feedback techniques to the keywords and keyphrases discovered by our system in finding other relevant documents from a corpus.

1 Introduction

Because they cannot divine the user's current work context, traditional information retrieval (IR) systems count on the user to play an active role in the formulation and refinement of search queries. However, in order for such systems to be effective, they often assume or require a level of technical and linguistic expertise far beyond what can be reasonably expected of the average user. *Proactive information retrieval* compensates for this limitation with the technique of relevance feedback, where the user is asked to manually mark up a representative set of search results as being relevant or irrelevant. The relevance feedback data is then used by the search engine to refine its results. However, explicit feedback systems are problematic in that users find them bothersome and time-consuming. A promising alternative approach is to gather *implicit* feedback by unobtrusively observing the user.

In this paper, we ask whether it is possible to determine the relevance of search result abstracts using a device which monitors the user's eyes while skimming. That is, can features such as gaze direction, duration, and pupil diameter tell us anything useful about how the user is processing and analyzing the text? If so, which of these features, or combination thereof, best correlate with the relevance of the text to the search query? And finally, how can the text passages

identified as relevant be used to improve search engine performance? To answer these questions, we present plans for an experiment to gather and evaluate the necessary data. We also present the results of a preliminary feasibility study.

2 Background

2.1 Eye Trackers

Devices to measure ocular indices have existed in various forms since the late 19th century [1], and the basic principles underlying today's corneal reflection eye trackers have not significantly changed since 1901 [2]. However, it is only with the advent of modern computer processing power that researchers have been able to fully realize the potential for this technology. A typical modern-day eye tracker consists of an infrared-sensitive video camera and an infrared LED, both directed towards one of the subject's eyes. The video images of the LED's reflection in the cornea are fed to a computer which measures pupil diameter and calculates the coordinates of the gaze point on a suitably positioned (perpendicularly planar) surface, such as a computer monitor [3, p. 60]. Samples are typically taken at high frequency (50 Hz and up) and then post-processed to identify *fixations* (spatially stable gazes) and *saccades* (rapid jumps).

Eye trackers must be calibrated before use, and in reading experiments typically need recalibration for every three screens of text read by the user. In order to obtain measurements precise enough to identify individual words, the on-screen text needs to be presented with generous line spacing at a font size of about 38 points. This corresponds to approximately twelve lines of text, or one paragraph, per screen. Furthermore, users' pupils tend to tire quickly, making precise measurements progressively more difficult. After about 20 minutes of reading, precision usually drops off rapidly.

2.2 Reading

It has long been known in neuropsychology that the retinal image is transmitted to the brain during fixations but not during saccades; therefore it is the fixations which represent the acquisition and processing of information. In normal reading, a reader does not fixate upon each word in sequence, but rather makes a rapid series of fixations followed by saccades which may skip over some words entirely. In addition, approximately 15% of all saccades occur backwards, to earlier text; this is known as a *regression*. Fixation times in normal reading range from 60 to 500 ms, with an average of about 250 ms [4]. Pupil dilation, regressions, and fixation length have all been used as measures of the mental work applied to process a given passage [5–7].

2.3 Previous Work

There is little previous research on using eye tracking in IR. Salojärvi et al. [8] used a task-focussed experiment where subjects were given a question and asked

to read through twelve newspaper headlines to find the answer. Eight of the headlines were irrelevant to the task and four were relevant; only one of the relevant headlines contained the answer. The experimenters transformed the raw eye-tracking data into 21 word-based features relating to the number and duration of fixations, fixations during regressions, pupil diameter, and saccade lengths. Various statistical information visualization methods, including standard self-organizing maps (SOMs), were then used to determine that the features were indeed sufficient to predict headline relevance with reasonable accuracy.

The only other study we are aware of is [9]. The setup is similar to [8], except that rather than having the subjects read headlines, they skim a page of Google search results. Also, the eye-tracking data is less precise; instead of measuring fixations on individual words, only three areas of interest are examined for each search result: the title, abstract, and metadata. Their analysis is currently underway so there are no published results yet on the relationship between ocular indices and document relevance.

Our study, named eFISK (Eye Fixations to Identify Salient Keywords) differs from the two previously cited in two ways. First, we gather ocular data on a per-word basis as the subjects skim through relatively long search result abstracts. Second, we also investigate not only whether certain indices correlate with relevance, but whether we can use this data to identify keywords which can serve as a new search query to discover other relevant documents.

3 Methodology

Our methodology will incorporate elements of [10] as their study is in a similar vein to ours. That study investigates the feasibility of using users' manual written annotations to identify salient passages, and compares this technique against standard binary relevance feedback on whole documents. Accordingly, we propose a task-based user study which will compare utility of standard relevance feedback techniques to the keywords and keyphrases discovered by eFISK in finding other relevant documents from a corpus.

Our experiment employs the German Indexing and Retrieval Test Data Base (GIRT), which is part of the Cross-Language Evaluation Forum (CLEF) corpus [11]. It consists of titles, abstracts, descriptor keywords, and other data on 151 319 German-language documents in the social sciences. Each document is associated with one or more descriptor keywords from the *Thesaurus for the Social Sciences* [12]; on average there are ten such descriptors per document.

We used the INQUERY [13, 14] search engine to build three data sets from GIRT. This was performed as follows: first, we selected at random one document from the GIRT corpus. We then used that document's abstract as an INQUERY search query, and examined the top ten search results (excluding the original document, which was always ranked #1). A particular search result was considered relevant if it had at least one descriptor keyword in common with the original document, and irrelevant if it had no descriptors in common. If the ten search results contained no less than three and no more than seven relevant

documents, then the top three relevant and top three irrelevant results, plus the original document, constitute a data set. Otherwise, the search was repeated with another randomly-selected source document.

Approximately twenty German-speaking subjects (mostly undergraduates at Saarland University) will be recruited to participate in the study. For each data set, each subject is instructed that he is to imagine himself a research assistant employed by an author in the social sciences. This author has read an abstract for a colleague's article (the source document of the data set) and has decided to write a book on the same topic. The subject's job is to find other documents that might be useful as background material for the author's book. The subject is given the colleague's abstract to read and is told that he will be presented with abstracts of six other documents returned by a search engine. He is to quickly skim through these search results, and, immediately after each one, to rate their relevance to the colleague's abstract on a five-point scale.

All seven abstracts will be presented on a computer monitor, with the six search result abstracts given in random order, and the subjects' eye movements will be recorded at 250 Hz by a head-mounted SMI EyeLink Model II eye tracker. The raw eye-tracking data will be mapped to fixations and saccades using software developed by Saarland University's Psycholinguistics Department. Fixation positions will be mapped to the nearest word (possibly disregarding a stop list of words with negligible semantic content). Fixations too distant from any word will be discarded as outliers. We will then compute various features for each individual word on the display; these features will include the 21 from [8] and possibly others of our own devising.

Our goal will then be to determine the following:

1. Do the subjective relevance ratings given by the subjects correlate with our purely objective criterion for relevance (i. e., common descriptors)?
2. Can relevance (using either the subject's measure or our objective measure) be reliably determined from the eye-tracking data?
3. If so, how can we extract keywords from the search results such that using them as a new or refined search query effictively returns further relevant documents?

The first of these questions can be answered with some basic statistics, and is one of the objects of the feasibility study presented later in this article. To answer the second question, we propose to follow [8] by employing statistical machine learning and visualization techniques such as SOMs.

The third question can be addressed with a simple experiment: The relevance ratings given by the subjects will be input as relevance feedback to the INQUERY search engine. (INQUERY is capabale of refining a search on the basis of explicit relevance feedback.) Let us refer to the refined search results thus obtained as R. Then, various schemes for identifying relevant keywords using the eye-tracking data will be tested, and the keywords thus extracted will form the basis of either a new search query or a refinement of the original one. Let us refer to the results of this query as S. We will then compare the residual precision and recall for R and S—that is, which set of results consistently returns the most documents

relevant to the stated topic. (As the subjects will no longer be available at this point, we will use our objective relevance measure.) Precision and recall will be assessed at cutoffs of 10 and 100 result documents, which represent, respectively, the user's initial overview of the search results, and the maximum number of results he might be expected to browse through interactively. The overall pattern of residual precision and recall scores will be analyzed statistically to quantify the extent to which the extracted keywords perform versus standard relevance feedback for finding further relevant documents in a collection.

4 Feasibility Study

Because eye-tracker time is costly, we conducted a feasibility study without an eye tracker to gather preliminary data for our project. The study was intended to answer the following questions:

- How many abstracts could a subject be expected to read in the 20-minute limit?
- How long will it take the subjects to perform relevance ratings?
- How reliable are the relevance ratings provided by the subjects?
- Will reading times be consistent across articles of the same length, or will subjects speed up or slow down as the experiment progresses?
- What other considerations, if any, do we need to take into account in the design of our experiment?

4.1 Methodology

The methodology of the feasibility study closely follows that described in Sect. 3. Not yet having access to the GIRT corpus, we randomly selected a topic currently receiving coverage in the news—in this case, the Iraqi Special Tribunal—and used the Google News Deutschland search engine to retrieve a list of candidate newspaper articles. Among the search results, we selected six articles, three of which we judged to be relevant to the topic, and three of which we judged to be irrelevant.

We then recruited fifteen human test subjects for a task-based evaluation. The participants were given the same instructions we have described for our main experiment, except that relevance ratings were made on a four-point rather than a five-point scale.

4.2 Results

In this section, we refer to the three relevant documents as *Saddam*, *Köpfe*, and *Theater*, and the three irrelevant documents as *Botschaft*, *Irak-Krieg*, and *Butler*.

Reading Time

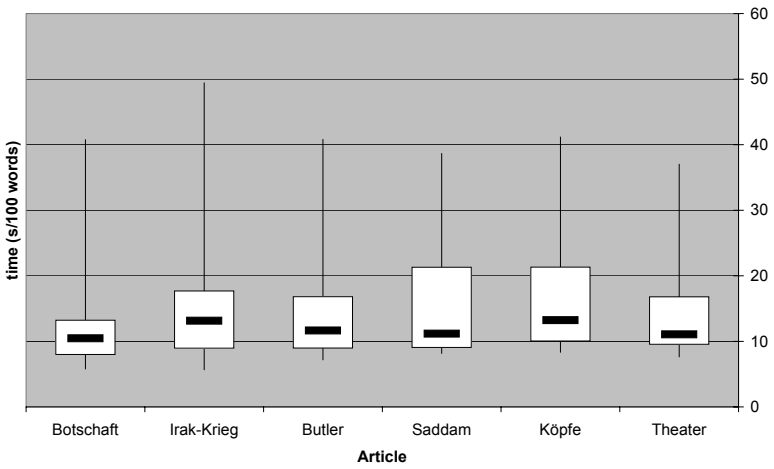

Fig. 1. Reading time per article

Total. Participation time for the entire experiment (instructions, six articles, and six questionnaires) ranged from 255 to 1035 s, with a mean time of 455 s ($\sigma = 195$ s).

Instructions. Reading time for the instruction text (322 words, including the topic description) was around two minutes ($\mu = 127\,\mathrm{s}, \sigma = 61\,\mathrm{s}$).

Rating Questionnaires. Each questionnaire had a mean completion time of 9 s ($\sigma = 4$ s); this includes time participants may have taken to rest before beginning the next article.

Articles. The six articles varied in length from 204 to 441 words. When normalized for length, median reading time per article was fairly consistent, ranging from 10.47 to 13.23 seconds per hundred words. Figure 1 summarizes the results in a box-and-whisker plot. (The plots in this report follow the standard convention, where the whiskers extend to the minimum and maximum values, the box extends to the first and third quartiles, and the median value divides the box.)

Because the articles were presented in a random order for each subject, we also plotted mean read time for each article in sequence. The results are summarized in Fig. 2. That is, the column labelled "1st" shows the reading time for the first article viewed by each subject, even though not every subject started with the same article. The plot, which is not normalized for article length, shows that

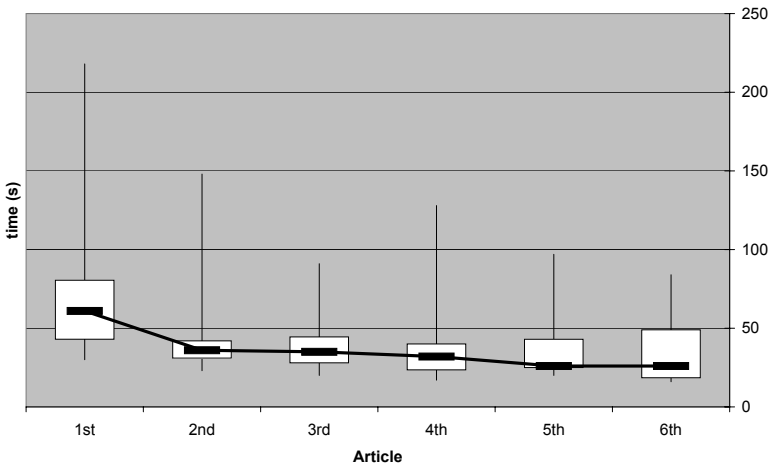

Fig. 2. Reading time per article in sequence

the median read time per article is monotonically non-increasing across time. Thus reading speed increases throughout the experiment irrespective of article length.

Relevance The human subjects' relevance assessments for each article were made on a four-point scale, where 0 represented "not at all relevant" and 3 "very relevant". As expected, relevance scores for the irrelevant articles were very low, whereas those for the relevant articles were much higher. Figure 3 summarizes our results.

Interjudge Agreement To compare interjudge agreement, we computed a correlation matrix for the relevance rankings. The results are summarized in the box-and-whisker plot of Fig. 4, which show the median Pearson correlation coefficient for each human subject. Agreement was moderate to high, with median r in the range $[0.3407, 0.7868]$.

4.3 Discussion

The highest median reading time per hundred words of article text was 13.23 seconds, so we can expect our subjects to skim through approximately 9000 words

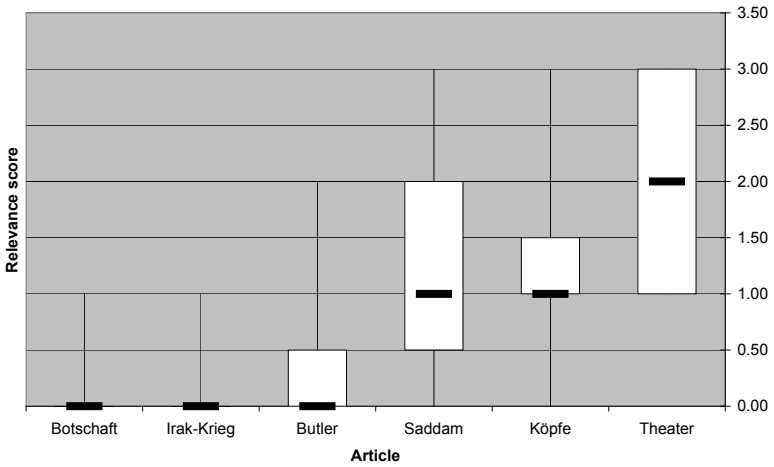

Fig. 3. Relevance

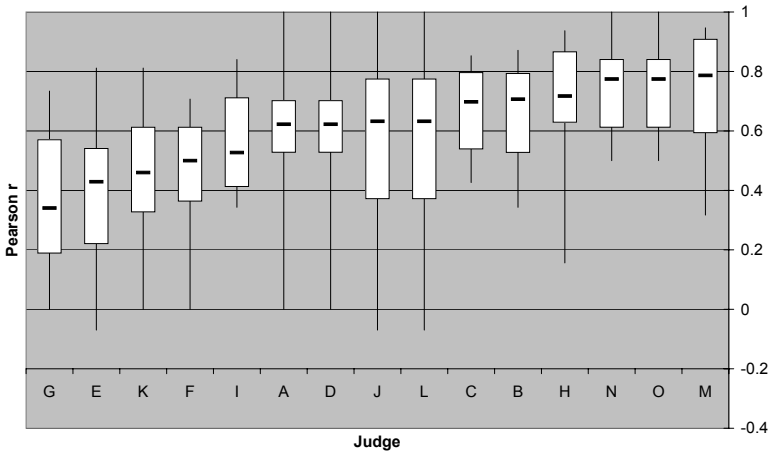

Fig. 4. Interjudge agreement

in 20 minutes. As the abstracts we will use for our final experiment are similar in length to the articles used in this preliminary experiment, this translates to about 30 abstracts per subject. With seven abstracts per data set, each subject could theoretically provide relevance ratings for up to four data sets. However, after accounting for the time involved in setting up, calibrating, and periodically recalibrating the eye tracker, we feel that three data sets per subject is a more realistic target.

The strong positive interjudge agreement is a good sign that subjects generally agree on what constitutes a relevant article. Furthermore, the subjects' subjective measurements of relevance correlated well with our "expert" classification. We feel that these results validate our task-based evaluation approach.

The data for reading time indicates that reading speed increases over time irrespective or article length. Therefore we may need to compensate for this in our final analysis of the ocular indices.

5 Conclusion

In this article, we have outlined plans for eFISK, an automated keyword extraction system using an eye tracker. The eye tracker unobtrusively monitors and records data about the user's gaze as he browses through a documents. The data is then analyzed to determine which words or passages the user found most relevant to his task. The passages thus extracted could be used as keywords to help index or annotate the document, or as input to a search engine to find similar documents. We proposed an evaluation wherein users would use an eye tracker while browsing through search engine results. We would then analyze this data to determine whether any ocular indices correlate with search result relevance. If there is a correlation (and based on the results of similar studies, we believe there will be), then we will investigate methods of extracting highly relevant keywords from the search results such that, when used as a search query, outperform the use of standard binary relevance feedback techniques. Finally, we described and presented the results of a feasibility study which essentially duplicates our proposed methodology, except without an eye tracker.

The results of our preliminary study confirmed the feasibility of our task-based evaluation approach. Because the amount of time a user can spend with an eye tracker is constrained both by physiological constraints and by our research budget, this "dry run" gave us data which will be helpful in economizing our use of the eye tracker. Finally, some elementary analyses on reading time showed us that we may need to adjust our results to account for certain factors, such as the order in which a given abstract is read.

With the feasibility study completed successfully, work has now begun on the main experiment. The raw eye-tracking data from our test subjects should be compiled by March 2005, and analysis will be performed in the months thereafter.

Acknoweldgement

This research is supported by the *Stiftung Rheinland–Pfalz für Innovation*, grant № 15202-386261/659. Thanks to Matthew Crocker, Alissa Melinger, and Andrea Weber for consultations respecting eye tracker usage.

References

1. Young, L.R., Sheena, D.: Survey of eye movement recording methods. Behavior Research Methods and Instrumentation **7** (1975) 397–439
2. Robinson, D.A.: The oculomotor control system: A review. Proceedings of the IEEE **56** (1968) 1032–1049
3. Duchowski, A.T.: Eye Tracking Methodology: Theory and Practice. Springer-Verlag, London (2003)
4. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin **124** (1998) 372–422
5. Beatty, J., Lucero-Wagoner, B.: The pupillary system. In Cacioppo, J.T., Tassinary, L.G., Berntson, G.G., eds.: The Handbook of Psychophysiology. Second edn. Cambridge University Press, Cambridge, UK (2000) 142–162
6. Reichle, E.D., Pollatsek, A., Fisher, D.L., Rayner, K.: Toward a model of eye movement control in reading. Psychological Review **105** (1998) 125–157
7. Engbert, R., Longtin, A., Kliegl, R.: A dynamical model of saccade generation in reading based on spatially distributed lexical processing. Vision Research **42** (2002) 621–636
8. Salojärvi, J., Kojo, I., Simola, J., Kaski, S.: Can relevance be inferred from eye movements in information retrieval? In: Proceedings of the Workshop on Self-Organizing Maps. (2003) 261–266
9. Granka, L.A., Joachims, T., Gay, G.: Eye-tracking analysis of user behavior in WWW search. In: SIGIR '04: Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, ACM Press (2004) 478–479
10. Golovchinsky, G., Price, M., Schilit, B.: 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, ACM Press (1999) 19–25
11. Kluck, M.: The GIRT data in the evaluation of CLIR systems – from 1997 until 2003. In Peters, C., Gonzalo, J., Braschler, M., Kluck, M., eds.: Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21–22, 2003, Revised Selected Papers. Volume 3237 of Lecture Notes in Computer Science. Springer-Verlag (2004) 376–390
12. Schott, H., ed.: Thesaurus Sozialwissenschaften – Thesaurus for the Social Sciences. Volume 1 & 2. Informationszentrum Sozialwissenschaften, Bonn (1999)
13. Broglio, J., Callan, J.P., Croft, W.B., Nachbar, D.W.: Document retrieval and routing using the INQUERY system. In Harman, D., ed.: Proceedings of the 3rd Text REtrieval Conference (TREC-3). NIST Special Publication 500-225 (1994) 22–29
14. Allan, J., Callan, J.P., Croft, W.B., Ballesteros, L., Byrd, D., Swan, R.C., Xu, J.: INQUERY does battle with TREC-6. In: Proceedings of the 6th Text REtrieval Conference (TREC-6). (1997) 169–206