# Measuring the Utility of Gaze Detection for Task Modeling: A Preliminary Study

Penelope Brooks[c], Khoo Yit Phang[c], Rachael Bradley[a], Douglas Oard,[a,b] Ryen White[b], François Guimbretière[c]
[a]College of Information Studies/ [b]UMIACS /[c]Department of Computer Science, University of Maryland, College Park, MD 20742
{pnoe,khooyp,rlb,oard,ryen,fguimbre}@umd.edu

## ABSTRACT

Search engines present readers with a list of documents ranked by predicted relevance to a keyword query. Salient sections of documents that are selected for examination can be highlighted using similar techniques. An ability to restructure information presentation based on an analyst's initial interactions with an information space might improve search outcomes. This study takes the first steps towards designing an analysis environment in which structural overlays evolve in response to an analyst's actions. Two sources of implicit feedback are explored: topical similarity to material included in a very brief written report, and eye behavior patterns. Results indicate that using eye-tracking can be as effective as lexical overlap, but more work is needed.

## Categories and Subject Descriptors

H.5.2 [**Information Systems**]: Information Interfaces and Presentation – *User Interfaces.*

## General Terms

Measurement, Design, Experimentation, Human Factors

## Keywords

Eye-tracking, implicit feedback, search, document summarization

## 1. INTRODUCTION

In the classical Information Retrieval (IR) paradigm, search systems rank documents in decreasing order of predicted topical relevance to a query. Using iterative query reformulation, searchers can often produce useful ranked lists, while brief query-focused summaries provide a basis for recognizing documents in those lists that are actually interesting. Selecting a link in the list can then reveal internal navigation cues to help users rapidly focus on the salient parts of a selected document. Modern IR systems accomplish these tasks based on lexical overlap with the query. While this can work well for carefully edited documents in which salient content exhibits strong locality, informal genre such as blogs or bulletin board postings frequently lack the explicit structure typically found in sources for which many present systems were designed. As informal sources become increasingly important, we will increasingly need to look beyond query-document term matching to identify observable behaviors that indicate an analyst's interests. In this paper, we focus on associating visual activity and report writing with a known (stimulated) interest.

Based on the increased interest in eye-tracking and the information that can be gathered by monitoring eye movements, we used an eye-tracker to monitor participants reading a document and creating a brief focused report. We hypothesized that eye-tracking would add useful evidence unavailable from lexical overlap with the content of the report. This approach was inspired by the *Inferring Relevance from Eye Movements Challenge 2005*, a multi-phase competition that seeks to advance the study of eye-movement data [3]. The *Challenge* recommends monitoring fixation duration, regression patterns and pupil size, while taking into account re-fixations and word skipping.

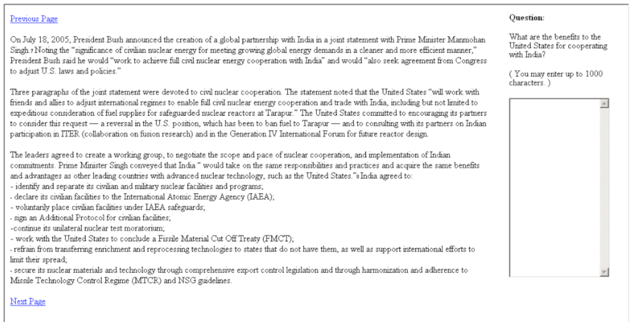
Our study builds upon research conducted in several fields, primarily eye-tracking, IR, psychology and linguistics. Vertegaal's work focuses on the challenges of designing an attentive interface by defining criteria that measure interest and addressing the discrepancy between visual interest and cognitive interest [9]. Eye-tracking researchers have monitored fixation, saccades (ballistic eye-movement from one target to another), and pupil size to indicate visual interest [1][2][5][6]. Granka *et al.* used eye-tracking to study visual fixation while using a search engine [1]. Our study differs from theirs in our focus on full documents rather than search-engine result lists. Puolmaki *et al.* conducted research with similar goals to ours, applying machine-learning to infer interest from reading behavior. They demonstrated a system that ranked documents in a manner consistent with the user's expectations [2]. We extend their work by measuring regression and pupil dilation, and by incorporating evidence from lexical overlap.

Our intent for this initial study was to focus on implicit feedback gathered by observing users' interaction with the system, associating those observations with content that is presented and reports that are created. Although our ultimate goal is to separately model interest and topical relevance, the design of a controlled user study requires a dependent variable that is comparable across study participants. For the study reported in this paper, we therefore stimulated a common interest by asking each participant to focus on a common set of topical questions while performing an IR task. We show that a combination of lexical overlap and eye tracking measures is always at least as useful, and sometimes more useful, than any single measure would be.

## 2. EXPERIMENT

Our experiment aims to determine if *interest*, an indirect variable representing the internal state of the reader, can be inferred from

*IUI '06 Workshop on Intelligent User Interfaces for Intelligence Analysis*, Jan 29, 2006, Sydney, Australia.



Previous Page

On July 18, 2005, President Bush announced the creation of a global partnership with India in a joint statement with Prime Minister Manmohan Singh.[7] Noting the “significance of civilian nuclear energy for meeting growing global energy demands in a cleaner and more efficient manner,” President Bush said he would “work to achieve full civil nuclear energy cooperation with India” and would “also seek agreement from Congress to adjust U.S. laws and policies.”

Three paragraphs of the joint statement were devoted to civil nuclear cooperation. The statement noted that the United States “will work with friends and allies to adjust international regimes to enable full civil nuclear energy cooperation and trade with India, including but not limited to expeditious consideration of fuel supplies for safeguarded nuclear reactors at Tarapur.” The United States committed to encouraging its partners to consider this request — a reversal in the U.S. position, which has been to ban fuel to Tarapur — and to consulting with its partners on Indian participation in ITER (collaboration on fusion research) and in the Generation IV International Forum for future reactor design.

The leaders agreed to create a working group, to negotiate the scope and pace of nuclear cooperation, and implementation of Indian commitments. Prime Minister Singh conveyed that India “ would take on the same responsibilities and practices and acquire the same benefits and advantages as other leading countries with advanced nuclear technology, such as the United States.”[8] India agreed to:
- identify and separate its civilian and military nuclear facilities and programs;
- declare its civilian facilities to the International Atomic Energy Agency (IAEA);
- voluntarily place civilian facilities under IAEA safeguards;
- sign an Additional Protocol for civilian facilities;
-continue its unilateral nuclear test moratorium;
- work with the United States to conclude a Fissile Material Cut Off Treaty (FMCT);
- refrain from transferring enrichment and reprocessing technologies to states that do not have them, as well as support international efforts to limit their spread;
- secure its nuclear materials and technology through comprehensive export control legislation and through harmonization and adherence to Missile Technology Control Regime (MTCR) and NSG guidelines.

Next Page

**Question:**

What are the benefits to the United States for cooperating with India?

( You may enter up to 1000 characters. )

**Figure 1. Screenshot of Trial.**

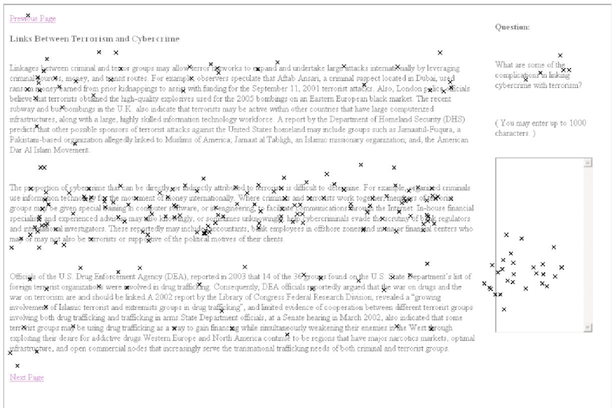
eye behavior. Based on previous work, we make the assumption that the amount of information processing, reflected by eye-movements during reading, correlates with visual interest [4]. Additionally, we assume that visual interest is equivalent to cognitive interest, as prior studies have suggested this to be true for relatively complex tasks such as in our study [3][9].

For this study, we designed an analytical reading task that required participants to answer a question with relevant information from a document. We define *(topical) relevance* as the information within a document needed to answer a question. By providing a question for the reader to answer, we controlled what they found relevant. We therefore rely on *relevance* as a surrogate for *interest* in this study. One question was designed for each document. To establish ground truth for relevance, each document was manually divided by the experimenters into sections – defined as a coherent group of words such as a paragraph, list item, or header. Five independent assessors then coded each section in the document as relevant or irrelevant with respect to the question.

Three Congressional Research Service (CRS) reports, two 7-screen documents (“Kuwait” and “Internet”) and one 18-screen document (“Cyberattack”), were read by each participant. CRS reports were selected because they are written in a standardized style that facilitated manual division into sections for ground truth relevance assessment, it is unlikely that participants would have read them, and they are long enough to ensure that both relevant and irrelevant information (to the question we posed) would be provided. These document lengths were selected based on a pilot study which indicated that reading multiple longer documents would result in fatigue.

For each document, one question was devised by one of the authors. The questions were open-ended, designed to be of moderate difficulty, and to draw on information from several portions of the document. For example, for the CRS report entitled *Terrorist Capabilities for Cyberattack: Overview and Policy Issues*, the question was: *“What are some of the complications in linking cybercrime with terrorism?”* In this example, “cybercrime” and “terrorism” appear frequently in the report, but “complication” does not appear at all and determining the answer requires synthesis across the whole document. Answer lengths between two and four sentences were expected.

We used an eye-tracker to measure the *number of fixations*, *fixation duration*, *number of regressions,* and *pupil size* of the participants' eyes. The results were then used as estimators of interest to rank sections. These measurement choices were motivated by previous work associating them to cognitive processes [4][7]. While those studies did not focus on reading, we speculate that the measures would also be useful for reading tasks. We hypothesized that we can generate useful list of relevant sections ranked by: higher number of fixations [H1]; longer fixation duration [H2]; larger pupil size [H3]; and greater number of regressions [H4]. We compare these interest-based ranking to a term-based ranking system and hypothesized that combining eye-tracking and lexical overlap would rank relevant sections more accurately than term-based ranking alone [H5].



Previous Page

Links Between Terrorism and Cybercrime

Linkages between criminal and terror groups may allow terror networks to expand and undertake large attacks internationally by leveraging criminal sources, money, and transit routes. For example, observers speculate that Aftab Ansari, a criminal suspect located in Dubai, used ransom money earned from prior kidnappings to assist with funding for the September 11, 2001 terrorist attacks. Also, London police officials believe that terrorists obtained the high-quality explosives used for the 2005 bombings on an Eastern European black market. The recent subway and bus bombings in the U.K. also indicate that terrorists may be active within other countries that have large computerized infrastructures, along with a large, highly skilled information technology workforce. A report by the Department of Homeland Security (DHS) predicts that other possible sponsors of terrorist attacks against the United States homeland may include groups such as Jamaatul-Fuqura, a Pakistani-based organization allegedly linked to Muslims of America; Jamaat al Tabligh, an Islamic missionary organization; and, the American Dar Al Islam Movement.

The proportion of cybercrime that can be directly or indirectly attributed to terrorists is difficult to determine. For example, organized criminals use information technology for the movement of money internationally. Where criminals and terrorists work together, members of terrorist groups may be given special training in computer software, or in engineering, to facilitate communications through the Internet. In-house financial specialists and experienced advisors may also knowingly, or sometimes unknowingly, help cybercriminals evade the scrutiny of bank regulators and international investigators. These reportedly may include accountants, bank employees in offshore zones and in major financial centers who may or may not also be terrorists or supportive of the political motives of their clients.

Officials of the U.S. Drug Enforcement Agency (DEA), reported in 2003 that 14 of the 36 groups found on the U.S. State Department's list of foreign terrorist organizations were involved in drug trafficking. Consequently, DEA officials reportedly argued that the war on drugs and the war on terrorism are and should be linked.A 2002 report by the Library of Congress Federal Research Division, revealed a “growing involvement of Islamic terrorist and extremists groups in drug trafficking”, and limited evidence of cooperation between different terrorist groups involving both drug trafficking and trafficking in arms.State Department officials, at a Senate hearing in March 2002, also indicated that some terrorist groups may be using drug trafficking as a way to gain financing while simultaneously weakening their enemies in the West through exploiting their desire for addictive drugs.Western Europe and North America continue to be regions that have major narcotics markets, optimal infrastructure, and open commercial nodes that increasingly serve the transnational trafficking needs of both criminal and terrorist groups.

Next Page

**Question:**

What are some of the complications in linking cybercrime with terrorism?

( You may enter up to 1000 characters. )

**Figure 2. Fixations overlaid onto sections**

We post-processed the eye-tracker data to determine the number of fixations, fixation duration, pupil size and number of regressions for each section. We measured *fixations*, when the eye “stops” for at least 60 milliseconds, in a window of 10x15 pixels with a time of 66 ms (determined by a pilot study). A *fixation location* may not be unique: we considered each fixation to the same section separately. For each fixation location, the *fixation duration* was measured. We defined a *regression* as a repeated fixation that involved eye movement to a section from another section or from the area on the screen where the answer was to be entered. We recorded the number of regressions that a participant made to each section. *Pupil size* (the average area of the pupil during a fixation) is provided by the eye-tracker in units of camera pixels. We recorded the average pupil size for each fixation location.

## 2.1 Protocol

We used a within-subjects design, allowing us to compare results across participants. Each participant experienced two levels of relevance (relevant and irrelevant) in each trial. To address order effects from document length and topic, we fully counterbalanced the order of the texts presentation. Eleven volunteers participated in our experiment, responding to an email sent out to university graduate and undergraduate students. Participants were not paid, but were provided with refreshments. Data from two participants were eliminated, one due to difficulty wearing the eye-tracker and the other due to excessive squinting. To maintain a fully counter balanced dataset, we therefore arbitrarily deleted data from three other participants for the preliminary analysis and reported on the final six participants in this paper.

Before the experiment, written instructions were given to each participant and read from a script by the experimenter. Each participant performed the experiment individually and began with a visual acuity test and reading speed assessment. The participant then dons the ISCAN ETL-500 head-mounted eye-tracker. Next, participants made seating adjustments while the experimenter ensured proper alignment of the eye-tracker optics.

After calibration, each participant was given a short practice trial, which served to clarify the instructions and to ensure comfort with the system. Next, the participant was presented with the first question and given an opportunity to clarify the question before proceeding to the timed trial. After indicating they were ready to proceed, the participant was presented with a document on the left side of the screen containing information pertinent to the question and the question and answer region on the right side of the screen (Figure 1).

Each participant was given five minutes to read the shorter documents and ten minutes to read the longer document. If an answer had not been completed when time expired, the document was removed from the screen and the participant was allowed to complete their answer. The time limit was intended to simulate the constraints in a real-world information acquisition task, forcing participants to read only as much as necessary, and limited their ability to pre-read and then answer the question from memory. At the end of the experiment, each participant was debriefed regarding the purpose of this research and comments were solicited. Additionally, they were asked to participate in a post-experiment questionnaire where their interest and knowledge of the topics presented was gauged.

The experiment was run on a 19" Dell LCD monitor, which had a native resolution of 1280x1024. The eye tracker was calibrated for each participant, both with the program provided with the eye-tracker and with an experiment-specific program. Minimizing calibration errors with our current head-mounted eye-tracker configuration proved to be challenging, in part due to discomfort and fatigue. We plan to use a Tobii desk-mounted system in the future, which may partially mitigate those effects.

## 3. DATA ANALYSIS

We used Lucene (available from http://lucene.apache.org) to compute a score for each section by indexing the sections as "documents" and then each participant's answer as a query. This produced a set of ranked lists, one for each participant. We then used *trec_eval* (from http://trec.nist.gov) to compute the Mean Uninterpolated Average Precision (MAP) for each measurement for each ground truth assessor. MAP is designed to model user satisfaction when traversing a ranked list. Given the output of the ranked list, Uninterpolated Average Precision is defined as:

$$AP = \frac{1}{R}\sum_{i=1}^{R}\left(\frac{i}{r_i}\right) \qquad \text{(Eq. 1)}$$

where R is the total number of relevant sections in the ranked list; $r_i$ is the rank of the relevant section *i*. The mean is then taken across participants, rather than across topics (as would be standard in a search engine evaluation).

We therefore obtained one MAP value for each assessor-document combination (a 4 by 3 matrix). Variations in document characteristics and assessor opinions make comparison of MAP across assessors and/or documents problematic with small sample sizes, so we restrict our analysis to comparison of alternative measurements with assessor and document held constant.

For the eye-tracking measurements, we computed a set of scores for each section by normalizing each measurement to fall within a [0,1] range such that a higher score represented a greater expected degree of relevance per our hypotheses. We then used the resulting scores to compute a set of ranked lists in which the elements in the list were sections and scored these ranked lists using *trec_eval* for each document-assessor pair.

In addition to fixations, fixation duration, pupil size, and regressions, three additional compound measures were also synthesized and used as the basis for constructing ranked lists. We observed during the experiment that participants were reading entire sections at the beginning of a document, but that they read less as they progressed through the document. To account for this observation, we computed a *fixation difference* measure, defined as the *number of fixations* minus the 7-section moving-average. The 7-section window was chosen to reduce the skew effect from short sections (e.g., titles and headers). The other two compound measures combine results from one eye-tracking measurement with the results from term-based search. We combined the scores by taking the maximum of *number of fixations* and *Lucene term-match score* for each section (MaxLucFix and MeanLucFix) in one case, and by replacing the maximum operator with the arithmetic mean of the two scores in the other case. Because these combinations were chosen *post hoc*, they should be treated only as suggestive of what might be achieved from evidence combination strategies.

Controlling for confounding variables posed some challenges. Differences in writing style or formatting could affect reading speed, so we chose documents from a single source. Another source of concern was prior knowledge by our participants. We had selected topics that we felt were likely to be new for the participants. Testing for *a priori knowledge* and *a priori interest* before the experiment would have been impractical, so we relied on post-experiment self-report data to screen for those factors. We used a post-experiment questionnaire asking each participant about their knowledge of and interest in the topics on a scale from 1 to 5. Responses to these questions were fairly consistent, so no data was excluded based on those factors. Another possible confounding variable was *reading speed*, since faster readers have been shown to make shorter fixations and fewer regressions [3].

## 4. RESULTS

Our preliminary analysis indicated that we may have given participants too little time to read the short "Kuwait" and "Internet" documents, so for this paper we have chosen to focus on the long "Cyberattack" document. For purposes of presentation, we have sorted the assessors in increasing order of the number of relevant sections that they identified in the "Cyberattack" document (2, 5, 7, 16, and 21 out of 83 sections, respectively). Based on this analysis we removed Assessor 1's scores, as that assessor only selected two relevant sections out of 83 sections; MAP values are overly sensitive to quantization noise when few relevant items are known. For each remaining assessor, MAP values averaged across the six-participant counterbalanced group are reported for each measure.

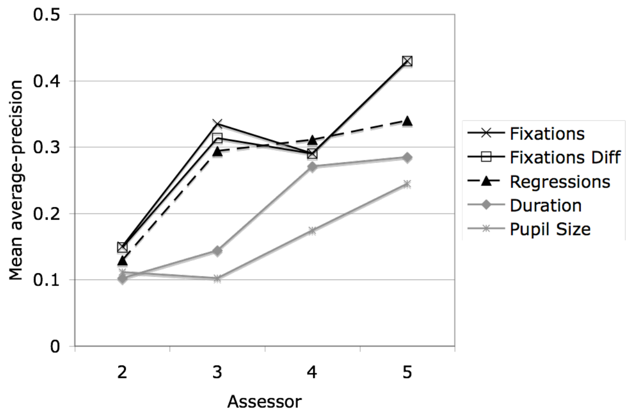


**Figure 3. Eye-tracking measurements for Cyberattack**

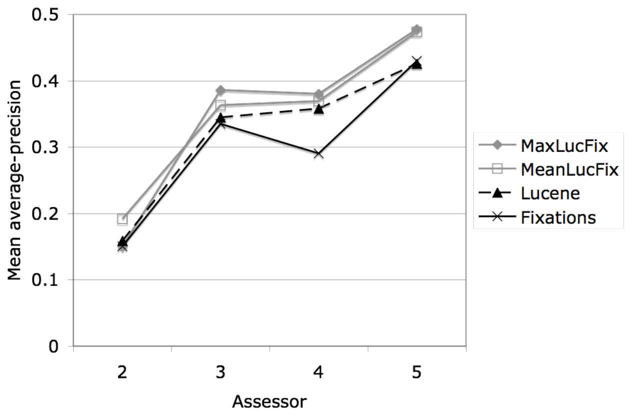


**Figure 4. Cyberattack Fixations, Lucene, Lucene+fixations**

Of the eye-tracking measurements, *number of fixations* (Fixations in Figure 3) results in the highest MAP score on average, followed by *number of regressions* (Regressions). *Fixation duration* (Duration) and *pupil size* (Pupil Size) yield much lower MAP scores. These observations tend to support hypotheses H1 and H4 more strongly than hypotheses H2 or H3. Surprisingly, *fixation difference* did not score better than *number of fixations*. We think that for relevant paragraphs, fixations increase more than it is being reduced by the position of the section in the document.

We speculate that Regressions may have turned out to be a better indicator of interest because of the way participants answered the question. Since they could answer as they were reading, they could frequently refer back to the relevant sections. Fixations would tend to increase as well due to re-reading. These factors suggest that we might want to treat Regressions from the answer area and Regressions from other sections separately in our analysis. This type of behavior would reasonably be expected to occur any time analysts are taking notes, so these two measurements may be particularly useful as predictors of interest.

Comparing Fixations with Lucene reveals little difference for three assessors and an apparent preference for Lucene for Assessor 4 (Figure 4). The composite scores MaxLucFix and MeanLucFix are at least 0.05 higher than Lucene alone for two of the four assessors, a criterion suggested by Sparck Jones as being noticeable to the user [7], and to be comparable to the better of Lucene or Fixations for the other two assessors. These results therefore tend to support hypothesis H5.

## 5. CONCLUSION

As with any initial foray into a new area, our results raise as many questions as they answer. Perhaps our most surprising result was that eye-tracking did as well as it did, yielding rankings that were typically nearly good as those obtained using term-matching. Combination of evidence seems to offer some promise, and we gained one insight (regarding Regressions from the answer area) that might help to design more nuanced evidence combination strategies. Several aspects of our study design, including the use of topical relevance as a surrogate for interest and the use of multiple independent assessors worked well, but our infelicitous choice of time limits for shorter documents and calibration difficulties with the eye-tracker we used will need to be addressed in future studies. Using well structured documents for these experiments facilitated creation of ground truth judgments, but if we are to achieve our ultimate goals we also need to begin to work with the informal genre that motivate this line of research.

## 6. ACKNOWLEDGMENTS

This work was partially supported by DARPA's GALE program.

## 7. REFERENCES

[1] Granka, L. A., Joachims, T., and Gay, G. 2004. Eye-tracking analysis of user behavior in WWW search. SIGIR '04.

[2] Puolamäki, K., Salojärvi, J., Savia, E., Simola, J., and Kaski, S. 2005. Combining eye movements and collaborative filtering for proactive information retrieval. SIGIR '05.

[3] Rayner, K. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. Psychological Bulletin 124(3), 1998, 372-422.

[4] Rudmann, D. S., McConkie, G. W., and Zheng, X. S. 2003. Speech and Gaze: Eye-tracking in cognitive state detection for HCI. ICMI '03.

[5] Salojärvi, J., Puolamäki, K., Simola, J., Kovanen, L., Kojo, I., and Kaski, S. (2005) Inferring Relevance from Eye Movements: Feature Extraction. Publications in Computer and Information Science, Report A82 (3 March 2005)

[6] Santella, A. and DeCarlo, D. 2004. Robust clustering of eye movement recordings for quantification of visual interest. ETRA'2004.

[7] Spärck Jones, K. Automatic indexing. Journal of Documentation, 30:393-432, 1974.

[8] Velichkovsky, B. M., Dornhoefer, S. M., Pannasch, S., and Unema, P. J. 2000. Visual fixations and level of attentional processing. ETRA '00.

[9] Vertegaal, R. 2002. Designing attentive interfaces. ETRA '02.