An Eye-Tracking Study of Query Reformulation

Carsten Eickhoff
Dept. of Computer Science
ETH Zurich, Switzerland
ecarsten@inf.ethz.ch

Sebastian Dungs
University of Duisburg-Essen
Duisburg, Germany
dungs@is.inf.uni-due.de

Vu Tran
University of Duisburg-Essen
Duisburg, Germany
vtran@is.inf.uni-due.de

ABSTRACT

Information about a user's domain knowledge and interest can be important signals for many information retrieval tasks such as query suggestion or result ranking. State-of-the-art user models rely on coarse-grained representations of the user's previous knowledge about a topic or domain. In this paper, we study query refinement using eye-tracking in order to gain precise and detailed insight into which terms the user was exposed to in a search session and which ones they showed a particular interest in. We measure fixations on the term level, allowing for a detailed model of user attention. To allow for a wide-spread exploitation of our findings, we generalize from the restrictive eye-gaze tracking to using more accessible signals: mouse cursor traces. Based on the public API of a popular search engine, we demonstrate how query suggestion candidates can be ranked according to traces of user attention and interest, resulting in significantly better performance than achieved by an attention-oblivious industry solution. Our experiments suggest that modelling term-level user attention can be achieved with great reliability and holds significant potential for supporting a range of traditional IR tasks.

Categories and Subject Descriptors

Information Systems [**Information Retrieval**]: Query Reformulation

Keywords

Eye-gaze Tracking; Knowledge Acquisition; Domain Expertise; Query Reformulation; Query Refinement; Query Suggestion; Mouse Cursor Tracking.

1. INTRODUCTION

Users of information retrieval systems have been shown to struggle with forming an accurate mental image of their information needs and the resources to satisfy them. Belkin et al. describe this observation as an *Anomalous State of Knowledge* (ASK), hindering users' query formulation and search success [4]. To mitigate this effect, Web search engines offer query suggestions and recommendations that guide the searcher towards popular queries, frequently issued by other users. It is, however, often unclear how relevant such suggestions are for the individual user, especially for non-transactional information needs. Ideally, we would like to promote those suggestions, that lead the user to relevant, novel and understandable documents rather than just generally popular ones. Personalized generation of query suggestions based on the user's previous search and interaction history has been found as one way to address this problem [10]. The proposed models, however, are coarse-grained and represent only high-level notions of the user's active query vocabulary. They consider, for example, all previously encountered terms (e.g., all terms present on recently visited Web sites) to be known and understandable. While this family of approaches makes a valuable first step towards integrating an understanding of the user's state of knowledge into the query suggestion process, one would require a system that can account for the user's vocabulary at a significantly finer granularity, ideally on the term level.

The same issue plays up at other points of the search process, for example during result ranking. State-of-the-art relevance models often include representations of the user and their specific context such as previous search history [48], preferences in terms of high-level topics [26], or content readability [11]. While such notions of text complexity have been demonstrated to significantly increase retrieval performance by providing users with resources of appropriate reading level, the readability metrics themselves are not personalized and rely on general complexity estimates based on a very diverse audience of users. Instead, it would be strongly desirable to know which exact terms the searcher is able to recognize, understand and actively use.

In this paper, we use eye-gaze fixations and cursor movement information in order to study which concrete terms the user is exposed to on Web pages and search engine result pages (SERPs) and how they are subsequently re-used as query terms. In this way, we make three novel contributions over the state of the art in user modelling. **(1)** For the first time, we inspect the evolution of users' active query vocabulary during Web search on the term-level in a qualitative user study. **(2)** Based on the eye-gaze signal, we model the likelihood of the searcher using a given term for query reformulation. **(3)** Eye-gaze tracking requires expensive hardware that would greatly restrict the exploitation and adaptation of our method. In order to make our in-

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767703.

sights flexibly applicable in most Web search settings, we substitute fixation information with traces of cursor movement.

2. RELATED WORK

Our investigation of related work will be guided by a number of topics that have been pursued in recent years. **(1)** First of all, we will revisit the body of work dedicated to measuring and tracking domain expertise, both statically as well as over time. **(2)** Secondly, there is an extensive line of work with the goal of query suggestion, reformulation and expansion. **(3)** And finally, we will give an overview of those eye-gaze or cursor-trace based studies that investigated user behaviour during the various stages of the search process.

Modern retrieval models rely on a diverse set of features in order to produce the final result ranking. One family of such features is concerned with measuring how familiar the searchers are with the topic of their information need. White et al. [49] investigated the behavior of domain experts and novices during Web search. They report higher likelihoods of search success for experts. Additionally, the authors observed gradual developments in domain expertise over the course of several weeks of search activity. Liu et al. [36] further find characteristic differences in user behaviour depending on the search task at hand. Wildemuth [50] studied concrete strategies and strategy types that domain experts and novices follow during information search. They find that over time, novices "learn" to use the same search patterns as experts if they are exposed to in-domain information for longer periods. This transition was studied in detail by Liu et al. [35], who investigated changes in domain expertise when searchers were following the same overarching tasks across multiple sessions. Eickhoff et al. [15] further showed session-level evidence of domain expertise increases in response to in-domain searches. The authors put a particular focus on the acquisition of new query terms which is explained by previous page visits. Zhang et al. [53] proposed an automatic prediction framework that was able to identify domain experts and novices based on a range of behavioral features. Kim et al. [29] cast the expertise problem as a combination of preferred reading level per topic. In this way, they tracked the notion of topic-normalized resource complexity for different users.

Query formulation can represent one of the cognitively most challenging steps in the information search process [13]. Building on the early investigations of Spink [47] and Saracevic [46], query suggestion functionality aims to aid the user at this initial stage. Chirita et al. [10] rely on information gathered on the user's local desktop in order to expand Web search queries. Kelly et al. [27] investigated query reformulation behaviour by offering query and term suggestions based on clustering and pseudo relevance feedback. They report a clear user preference for query suggestions over term suggestions. Gao et al. [16] rely on information mined from large-scale query log files to provide suggestions. Independently, Song and He [45], as well as Ma et al. [38], propose personalized query suggestions by analysing the user's previous click behaviour within the session in order to mine suggestion terms from skipped and visited documents.

Eye-gaze tracking has been used for unobtrusive tracking of user attention and interest for several decades [24, 43]. In the information retrieval community, a wide array of studies and applications have been proposed in recent years. Cutrell and Guan [12] compare various degrees of SERP verbosity for different task types, finding that inherently navigational tasks require less information per item than informational ones. Granka et al. [17, 23] conducted an eye-tracking study of users' interaction with search result lists. They confirm several established notions such as the well-known position bias of user attention and note significant potential for using eye gaze signals as implicit relevance indicators. In a number of small-scale (5-8 participants) qualitative studies, Kunze et al. use head-mounted eye gaze tracking devices for inferring language expertise [30] and document types [31] based on reading styles and eye movement patterns. Williams and Morris [51] contrast the fixation duration of familiar and unknown words during silent reading. They find that unfamiliar words receive significantly longer attention windows than known ones. We will revisit this finding in Section 6 of this paper.

Salojärvi et al. [44], as well as Brooks et al. [6] investigate a range of low-level eye-gaze features for inferring passage-level relevance labels for information retrieval and collaborative filtering[42]. Loboda et al. [37] investigated eye-gaze indicators of sentence-level relevance. They found the overall number of fixations, the number of first pass fixations as well as the total viewing time to carry most indicative power. Interestingly, and somewhat conflicting with the findings of both this paper as well as [8], the authors could not find any clues of term-level relevance. Buscher et al. [8, 7] use fixation length and frequency as relevance indicators for query expansion and search result personalization. Their work is closely related to the application scenario presented in Section 6 of this paper. While their approach relied on shallow term-level feedback mechanisms, we leverage semantic information via related or synonymous terms as well as using a wider array of eye-gaze signals. Ajanki et al. [2, 41] use eye tracking hardware to infer document-level relevance across a manually curated document corpus and, subsequently, generate alternate queries based on salient terms. There is some evidence of the origin of previously encountered Terms being reused in active vocabulary [51, 15], but there has not yet been a dedicated study to further investigate and understand this vocabulary acquisition process.

Independently, Guo and Agichtein [19] as well as Huang et al. [22] propose to infer user eye gaze and, implicitly, user interest, from mouse cursor position and movement. The authors show a substantial overlap between both sources that motivates further exploitation as for example presented later in this work. In a follow-up publication [18], the authors exploit a similar range of signals in order to infer post-click document relevance. Finally, Huang et al. [21] predict click-through on the basis of cursor position and movement signals.

Our work differs from the above body of research in that: **(1)** it measures user vocabulary knowledge and its development over time at a much finer granularity than previous efforts, which mainly concentrate on broad topic verticals. **(2)** Our model is based on actual observations of user attention to individual terms as evidenced by eye-gaze and mouse-cursor traces. Previous work on domain expertise is based mainly on the posterior analysis of search engine log files in which the mere presence or absence of a term on a page or SERP is regarded as a signal. Information about whether the user actually saw the term, and if so, how long the engagement lasted, are not available. **(3)** The breadth

of existing eye-gaze and mouse cursor movement studies has investigated many aspects of the search process. However, an in-depth study of query reformulation behaviour at the term level, as we propose in this paper, has not yet been attempted. The timeliness and relevance of this study is further evidenced by several pieces of prior related work that explicitly state the need for a more qualitative understanding of query reformulation and term acquisition on the Web (e.g., [15]).

3. METHODOLOGY

The user study took place in a controlled lab environment at a university campus. Participants were recruited through advertisements (flyers and Internet ads) and public announcements. Overall, 17 persons (8 female and 9 male) participated in our study. All participants were students majoring in a range of different, often IT-related, subjects. All participants were between 19 and 27 year old (average 22.7). On average, each participant had 9.8 years of active Internet usage. All participants had experience with Web search engines as well as searching in digital libraries. Experiments lasted for about 60 minutes and participants were compensated by a payment of € 10. 7 (3 female, 4 male) of the participants were part of an initial pilot study while the remaining 10 persons contributed to the final experiments that will be analysed in the further course of this paper. Due to incremental changes to the experimental setup as well as occasional technical glitches during the pilot study, we report only the outcomes of the final experiment.

At the beginning of each experiment, a short introduction to the study, including the eye tracking hardware, was given. Participants were not informed about the concrete research questions and hypotheses of the study. Subsequently, the eye tracking hardware was calibrated and participants were told to maintain a firm yet comfortable seating position for the duration of the experiment.

All sessions were conducted on a Windows 7 system with 22" (1680 × 1150) display, running a Firefox 25.0.1 Web browser. Eye-gaze traces were recorded with an SMI RED remote eye tracking system that is integrated into the monitor. This setup is considered to be less intrusive than head-mounted alternatives. The system captures gaze positions at an update frequency of 60Hz and an accuracy of 0.4°. The recordings and analysis were made using iViewX, Experiment Center, and Begaze 3.4. We rely on a number of essential smoothing techniques such as fixation grouping, e.g., described in [5]. Our instrumented Web browser saves screenshots of each accessed page and the final mapping between fixation coordinates and terms rendered on screen is established via OCR technology. This approach has the advantage of making **all** rendered text accessible, regardless whether it was expressed in plain HTML, or encapsulated in AJAX or JS containers. Previous work [14] found that the inability to parse text contained in such elements can significantly limit the performance of analytical and inference methods.

After the calibration, the Web browser was used to present the questionnaires and tasks to the users. The same tab was used for the questionnaires as well as the input boxes for task completion. Participants were asked to leave the instruction tab open at all times and to use other tabs to their liking. Task presentation and questionnaires were structured as follows:

To ensure task diversity yet obtain a reasonable amount of overlap between tasks, we hand-picked 6 topics (ids 31, 38, 41, 42, 55, 69) from the 2006 and 2007 editions of the TREC Question Answering Track's complex interactive QA task [28]. In the initial pilot study, we found that scientific and biomedical tasks resulted in more frequent query reformulations per session as users gradually acquired the relevant domain vocabulary. Many of the politically and societally motivated tasks have been comprehensively investigated and summarized, due to the time that has passed since the original formulation of the tasks. This resulted in socio-political tasks being mostly answerable with a single page visit. As a consequence, we expanded our pool by several additional topics (ids 53, 70, 72, 73). Our final selection consists of 10 tasks, originating mainly from the bio-medical domain.

Each participant was asked to complete three search tasks. For each task, they were offered two options and were told to choose a task according to their personal preference. As a result, we obtain a total of 30 search sessions (10 participants each choosing 3 tasks). We decided to offer this choice to allow participants to focus on tasks that they found personally interesting, which in turn is expected to spark better engagement and richer interactions during the session. A Web application scheduled the tasks, ensuring that each of the 10 topics was offered for selection equally often and in non-repeating pairings (in our case each topic was shown exactly 6 times to offer 10 participants each 3 choices of two of the 10 available topics). After the task selection, a demographic questionnaire was given to the participants. Subsequently, each of the actual tasks was presented in the browser and accompanied by short pre-task and post-task questionnaires. Pre-task questionnaires covered topic specific aspects like familiarity and confidence as well as perceived difficulty. Post-task questionnaires again covered the aspects of perceived task difficulty as well as perceived completeness and quality of the answers given by the users. We allocated up to 20 minutes time per search task, resulting in an overall duration of up to 60 minutes per participant across all tasks. After all three working tasks had been completed, the searcher's general opinion of the experiment was elicited in a post-experiment questionnaire.

Table 1 discusses some of the salient characteristics in anticipated and actual task difficulty according to the pre- and post session questionnaires. Since we offered the participants to choose their tasks, we can note interesting differences in task frequency. We report averages across all participants that selected and completed a given task. Answers were given on a 5-point scale ranging between settings of 1 (not at all [easy, familiar, ...]) to 5 (very [easy, familiar, ...]). We can note distinct task-specific levels of difficulty originating from different coverage of the topic on the Web. Due to the mostly bio-medical nature of the tasks and the lack of a relevant formal background knowledge among the participants, we see relatively low scores of prior familiarity. This can be further seen by the fact that our bio-medical laypeople generally overestimated the difficulty of the task (with the exception of Task 31, which also showed the lowest overall participant satisfaction with the results of their search activity).

Table 1: Lab-study task characteristics.

Task Id	Frequency	familiar (pre)	interesting (pre)	easy (pre)	easy (post)	understandable (post)	found good results (post)
31	4	2.25	3.75	2.50	1.25	3.50	1.75
38	4	2.75	3.50	3.00	2.75	4.50	2.75
41	1	4.00	4.00	2.00	5.00	5.00	5.00
42	4	1.00	2.75	2.50	4.00	4.50	3.25
53	3	2.33	3.67	2.67	3.33	4.00	3.33
55	5	1.20	2.80	3.40	3.20	4.60	3.40
69	1	1.00	4.00	4.00	5.00	5.00	4.00
70	6	1.17	2.5	2.5	2.33	3.5	2.67
72	0	-	-	-	-	-	-
73	2	1.50	4.50	3.50	4.00	4.50	4.50
Overall	30	1.73	3.23	2.83	3.00	4.17	3.06

4. EYE-TRACKING EXPERIMENTS

In this section, we present the outcome of the previously described eye-tracking study. Following previous work [43], our experiments centrally consider eye-gaze *fixations*, brief periods of time when the reader focuses on a single location, during which no significant eye movement can be noted. The frequency and duration for which the gaze is kept steady are established signals of user attention.

4.1 Literal Term Acquisition

As a starting point to our investigation, let us revisit the findings of Eickhoff et al. [15] who conducted a log-based analysis of query term acquisition. They report that a significant share of all subsequently added query terms in a search session were present on SERPs and previously visited pages earlier in the same session. The authors interpret this observation as evidence of query term acquisition, but already state that, based solely on log files, there is no reliable way of determining which of these co-occurrences are genuine (i.e., the user actually sees and reuses a new term). Intuition suggests that many such cases are due to chance and are never really seen, processed and acquired by the user, e.g., because they were outside of the visible screen area displayed to the reader. To correct for these inaccuracies we reproduce their log-based approach for the search sessions collected in the previous section and contrast it with actual eye-gaze fixations. Similar to previous research, we find a share of 43% of all added query terms to have occurred on previously visited pages and SERPs. The number of actually fixated terms that later on are being used as query terms, however, is much lower (21%). As surmised originally, mere term presence is too coarse an estimator of query vocabulary evolution. The attention-based subset that we capture via eye-gaze tracking, instead, describes what the user has actually seen and potentially adopted from SERPs and Web pages.

To begin our in-depth investigation of term-level user attention and its effect on query reformulation behaviour, let us briefly introduce some necessary notation. Each search session comprises a number of SERPs and visited pages. We break these pages down into white space-separated tokens t. We distinguish between those tokens that appear in any of the session's queries T_q and the much larger remaining set of non-query terms T_n. The overall set of all tokens displayed in the session is given by the union $T = T_q \cup T_n$. Besides the displayed tokens, we also collect F, the set of all eye gaze fixations that were measured in the session. Each fixation $f \in F$ is described in terms of its duration $dur(f)$ and screen coordinate $loc(f)$. Using the previously described mapping between screen coordinates and display terms, we can now associate fixations with the terms that were rendered at the respective coordinates on the screen. As for display tokens, we can now subdivide fixations into those that rest on query terms (F_q) and those that focus any other terms (F_n). For this study, we disregard any fixations that, after application of a tolerance threshold of 5 pixels, do not coincide with the bounding box coordinates of a display token. This step removes all fixations that fall on browser control elements, images or page margins.

Table 2: Term-level fixation statistics show higher-than-average user attention on query terms.

	Non-query terms	Query terms
Relative frequency on page	0.83	0.17
Share of overall fixations	0.68	0.32
Rel. number of fixations per token	0.25	**0.57***
Share of overall fixation duration	0.61	0.39
Avg. fixation duration per token	58ms	**218ms***

$$att_{rel}(T_q) = \frac{|F_q|}{|T_q|} \tag{1}$$

$$dur_{rel}(F_q) = \frac{\sum_{f \in F_q} dur(f)}{\sum_{f \in F} dur(f)} \tag{2}$$

Let us now compare the way in which users interact with query terms and non-query terms. Table 2 shows an overview of several eye-gaze fixation statistics. We note that the general distribution of tokens in T is heavily biased towards non-query terms. Unsurprisingly, as a consequence of this skewed distribution, T_n also receives the largest share of the session's fixations. If we, however, discount those absolute fixation frequencies by the overall distribution of query terms and non-query terms (see Equation 1), it becomes apparent that tokens in T_q receive a significantly higher relative number of fixations per token than their non-query counterparts T_n. Similarly, we notice that the duration of individual fixations, both in terms of the absolute per-fixation duration $dur(f)$ as well as the relative share of the overall fixation duration $dur_{rel}(F_q)$, are biased towards giving significantly more attention to T_q. The statistical significance of the differences between query and non-query terms was determined by means of a Wilcoxon signed-rank test at $\alpha = 0.05$ level.

In practice, these effects result in situations as observed for example in Session 001, during which a participant with-

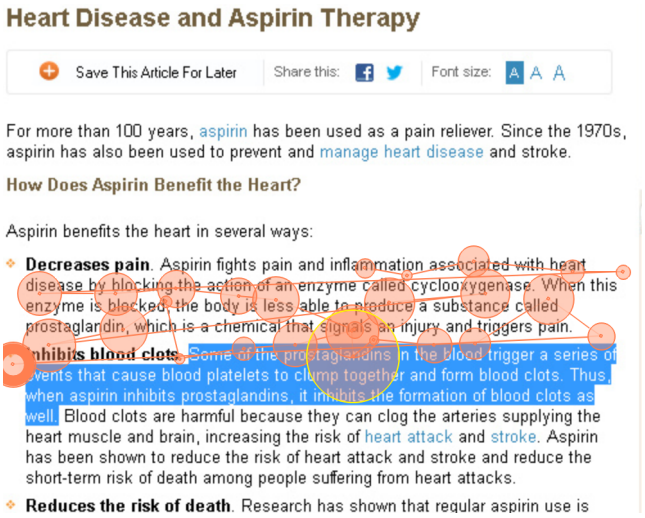

Figure 1: Eye-gaze patterns of Session 001 prior to query reformulation show strong evidence for acquisition of the medical term "prostaglandin" occurring in a paragraph of user-highlighted text.

out formal medical background was solving Task 42: "What effect does Aspirin have on coronary heart diseases?". The participant started with a query that consisted of all nouns taken from the original task descriptor and studied a number of high-ranked results. On one of the visited pages, the participant encountered a text passage discussing the interplay between Aspirin and a particular kind of lipid compounds, so-called *prostaglandins*. After having read the paragraph, the participant reformulated the query, adding prostaglandin as a new query term. Figure 1 shows a visual representation of the eye-tracker output for this acquisition of medical jargon. The orange lines display gaze patterns and the size of circles depicts the duration of each fixation. This session, as well as many similar examples that we encountered in our experiments, motivate the use of term-level eye-gaze tracking output for modelling user attention. In Section 6, we will demonstrate that exact scenario at the example of re-ranking query suggestion candidates.

4.2 Semantic Proximity of Reformulations

Up to this point, we studied fixated words on SERPs and visited pages that were subsequently picked up as query terms for reformulation. While those literal term acquisitions occur frequently, the majority of reformulations cannot be explained in this way. We will now, instead, inspect the semantic relatedness between reformulation terms and previously fixated ones. The updated hypothesis being that even though eye-gaze fixations do not literally forecast all newly added terms, they describe the user's interest accurately enough to allow for us to infer which semantic cluster of terms will indeed be used. This scenario includes cases such as synonyms or antonyms of fixated terms being employed for reformulation. To measure semantic proximity, we rely on WordNet and the well-known Leacock-Chodorow similarity [34]. The LCH metric is based on the *length* of the shortest path between the synsets containing the two terms and the maximum taxonomy depth

Table 3: Per-term fixation likelihood and duration show a general upwards trend as the semantic proximity to query terms increases.

	fixation duration	fixation likelihood
$LCH < 0.25$	30ms	0.09
$0.25 \leq LCH < 0.5$	36ms	0.15
$0.5 \leq LCH < 0.75$	39ms	0.23
$0.75 \leq LCH < 1.0$	34ms	0.18
$1.0 \leq LCH < 1.25$	32ms	0.13
$1.25 \leq LCH < 1.5$	38ms	0.27
$1.5 \leq LCH < 1.75$	**45ms***	**0.35***
$1.75 \leq LCH < 2.0$	**53ms***	**0.43***
$LCH \geq 2.0$	**86ms***	**0.49***

D_t. We used the WS4J implementation of LCH (`https://code.google.com/p/ws4j/`).

$$LCH = -log(\frac{length}{2 \times D_t}) \quad (3)$$

To test our hypothesis, we will inspect the distribution of user attention across the spectrum of LCH similarity scores between query terms and fixated terms. Please note that for this experiment, we exclude all direct query term occurrences from the comparison to allow for an exclusive and unperturbed study of semantically related terms. Table 3 shows the per-term likelihood and duration of fixation as functions of semantic proximity in terms of LCH scores. Also note that the LCH scale, in this case, does not reach its full extent (normally around scores of 3.0) because we pruned away all literal query term occurrences for this experiment. We can observe an initial local attention peak at an early point of the scale (LCH scores between 0.5 and 0.75) which is due to the highly frequent, but at the same time hardly related, stop words. As we, however, approach the far end of the LCH scale, we note a significant increase in user attention, bearing evidence of the importance of semantic proximity, even if literal term overlap is not given. For the highest proximity ranges ($LCH \geq 1.5$), we note a statistically significant increase in both fixation duration and likelihood as compared to each of the lower ranges ($LCH < 1.5$). Statistical significance of improvements was measured by means of a Wilcoxon signed rank test at $\alpha < 0.05$-level. Later on, in Section 6 of this paper, we will make use of this observation for model smoothing purposes.

We experimented with a number of alternative measures of semantic proximity including HSO [20], LESK [3], or WUP [52]. All considered metrics show the same initial rise of fixation frequency and duration followed by a monotonic rise as semantic proximity scores increased. There did not seem to be a systematically beneficial choice of metrics. Finally, LCH was chosen due to its low computational complexity.

4.3 Term Length and Complexity

In the previous sections, we showed how future query terms as well as semantically related terms received significantly higher-than-average amounts of user attention. In order to verify that this conclusion is indeed valid and not just due to hidden correlations with other unobserved effects, the

Table 4: Query terms receive significantly longer fixations than unrelated terms of comparable length.

	all terms		stop words removed	
Length	query terms	non-query terms	query terms	non-query terms
< 4	30ms	29ms	32ms	30ms
4-7	**38ms***	30ms	**39ms***	32ms
8-11	**49ms***	33ms	**49ms***	33ms
12-15	**55ms***	38ms	**55ms***	38ms
16-19	**69ms***	41ms	**69ms***	41ms
≥ 20	**82ms***	51ms	**82ms***	51ms

Table 5: Query terms receive significantly longer per-term fixations than unrelated terms of comparable complexity.

	all terms		stop words removed	
AOA	query terms	non-query terms	query terms	non-query terms
< 4	30ms	31ms	31ms	32ms
4-6	35ms	32ms	35ms	33ms
6-8	**49ms***	32ms	**49ms***	34ms
8-10	**72ms***	36ms	**72ms***	36ms
10-12	**112ms***	41ms	**112ms***	41ms
12-14	**173ms***	48ms	**173ms***	48ms
14-16	**228ms***	59ms	**228ms***	59ms
16-18	**287ms***	64ms	**287ms***	64ms
≥ 18	**384ms***	89ms	**384ms***	89ms

following paragraphs will discuss the effects of term length, complexity and stop words on user attention.

It is intuitively plausible that term length should play a role in the division of user attention. Longer terms take longer to read and have a greater likelihood of capturing chance fixations. Table 4 shows an overview of fixation duration on terms of various lengths. We note a monotonic, yet mild increase in fixation duration as terms grow longer. This applies to both query terms and non-query terms alike. However, for all but the very shortest length category, we see significantly longer durations for query terms than for arbitrary ones. Statistical significance of improvements was measured by means of a Wilcoxon signed rank test at $\alpha < 0.05$-level.

Previous work established how complex or unknown terms generally captivate the reader's attention longer than easy or well-known ones. Given the mainly scientific domain of our tasks, it is possible that the increase in attention, that we observed previously, is explained by the terms inherent complexity rather than their relevance. To investigate this hypothesis, we rely on previous work by Kuperman et al. [32], who compiled a list of 50,000 English terms along with their average age of acquisition (AOA). Common words such as *"the"*, *"a"*, or *"house"* show low ages of acquisition, usually between 1.6 and 4.0. Higher ages of acquisition, on the other hand, indicate greater term specificity and complexity, such as *"epithalamium"* with an average AOA of 17.67. Table 5 shows the distribution of attention received by query terms and non-query terms of given AOA grades. For both classes, we can note a mild upwards trend as AOA ranks increase. A much greater margin, however, separates the two classes from each other. We can therefore safely assume, that while term complexity plays a role in the distribution of user attention among terms on the screen, the governing factor is indeed topical relevance, as assumed in Sections 4.1 and 4.2.

As a conclusion to our overview of potential hidden factors correlated with user attention, we would like to draw special attention to stop words. This class of highly frequent but individually uninformative terms form the "syntactic glue" that ties together the content terms that carry actual meaning. Using the popular *Snowball* list of stop words [40], we observe that stop words receive much less attention (on average 30 ms per term) than non stop words (65 ms). In order to control for the influence of stop words on the previous investigations of term complexity and length, both Tables 4 and 5 show the respective measurements after stop words were removed. Following intuition, stop words are short (average length of 3.9 characters) and non-complex (average AOA of 5.09). Accordingly, we exclusively observe changes in the early rows that list short and of low complexity. Even in those categories, the changes are only marginal, supporting the conclusion that stop words do not play a special role with respect to user attention but rather follow the general trend dictated by their individual lengths and complexities.

Table 6: The effect of individual query reformulation strategies on LCH proximity between queries and user attention.

Strategy	Observed Frequency	Avg. change in $LCH(t)$
Specification	48%	+24%
Generalization	16%	+18%
Reformulation	36%	+37%

4.4 Alternative Reformulation Strategies

Throughout this section, we investigate various ways of explaining query reformulation in which users add new terms to an existing query. Lau and Horvitz [33] refer to this case as *specialization* since the focus of the conjunctive query is narrowed down with the addition of each new query term. They introduce 2 additional types of modifications, *generalization*, during which terms are removed which results in a broader, more diverse set of results, and *reformulation*, the exchange of one query term for another. This final case can effectively result in a radical topic shift depending on how semantically similar the terms are. To investigate the effect of all three major reformulation strategies, we group all instances of query reformulations depending on whether terms were added (specialization), removed (generalization) or exchanged (reformulation) and measure the average semantic similarity between the terms fixated by the user and the query, before and after the reformulation took place and report the relative difference. Table 6 shows the results of this experiment. As we can see, all three reformulation strategies result in a net gain in LCH scores, increasing the similarity between fixated terms and produced query vocabulary. The highest individual gains were noted for reformulations. Here, typically, mildly related terms are exchanged for highly related ones. Generalizations introduce the least dramatic increase in similarity. Mostly, this strategy corrects for previously added terms that drove the result set in an undesired topical direction.

5. CURSOR-TRACKING EXPERIMENTS

Based on the previously studied eye-gaze signals, we were able to show various forms of evidence of term acquisition during Web search. The main shortcoming of this approach (e.g., for improving the ranking quality) is the need for expensive, non-portable hardware to collect the required eye-gaze traces. As a consequence, the number of observations is very limited. Previous work has found a strong correlation

Table 7: Term-level cursor hover statistics.

	Non-query terms	Query terms
Relative frequency	0.79	0.21
Share of overall hovers	0.74	0.26
Rel. number of hovers per token	0.045	**0.059***
Share of overall hover duration	0.78	0.22
Avg. hover duration per token	130ms	**135ms***

between eye gaze and cursor movements [19, 22]. In this section, we roll out the previously described experimental setup, identically, to a large and diverse audience of crowdsourcing workers on Amazon Mechanical Turk (AMT). Since the availability of eye tracking hardware cannot be ensured in this altered setup, we will instead rely on traces of cursor movement to infer user attention. Where, previously, we investigated duration and frequency of fixations, we will now replace them with the duration and frequency of mouse cursor hovers over terms on the screen. Aside from this substitution, all previously introduced formulas and notations remain the same.

We conducted an initial experiment comprising 500 search sessions. 137 individual workers were presented with a randomly selected topic out of the previously introduced pool of 10 and were redirected to a custom-built Web search engine based on the public API of a popular search engine provider. Each search session was remunerated with $ 0.25. The crowd demographics are more diverse in terms of age, level of education, field of study, and language background. The gender split is comparable to the situation described earlier for the lab study. Table 7 shows cursor hover statistics across 500 individual search sessions. Again, the majority of tokens on the page does not fall into the query term category T_q. Despite this heavily skewed prior distribution, the relative number of hovers per token as well as the time for which the cursor rests on query terms is significantly greater than for non-query terms.

Let us now move on from literal query term occurrences to a final inspection of semantic relatedness. Table 8 compares per-term hover duration and likelihood for varying ranges of LCH proximity scores. While the middle ground of the distribution is less indicative than in the fixation case earlier, we still note a significant increase in hover likelihood and duration between terms of low to moderate semantic proximity ($LCH < 1.0$) and those of high semantic proximity ($LCH > 1.75$). Statistical significance of improvements was measured by means of a Wilcoxon signed rank test at $\alpha < 0.05$-level. Short hovers are much more likely to be caused by unrelated terms while very long hovers occur more frequently for related terms. This finding is supported by recent work by Ageev et al. [1], who investigate the connection between mouse cursor hover durations over relevant document passages to the results of user generated document summaries. Especially, their Figure 3 is in line with our findings here as well as earlier in Table 3.

When comparing our findings to the ones presented earlier in Section 4, it becomes obvious, that fixations are richer and more accurate predictors of user attention than cursor traces. The majority of users only occasionally use the mouse cursor in order to highlight text, mark their current reading position or follow textual hyper links. The result are the previously discussed correlations with topically relevant terms. For a share of 13.6% of the 500 search sessions, however, we observed a stronger connection between eye gaze and mouse movement. Here, users employed the cursor to trace every line of text as they read, creating a pattern that closely mimics the shape of eye gaze traces. Informal discussion with industry researchers from Google revealed that they as well noted this behaviour for 12 - 15% of their user base during an earlier, yet unpublished, large-scale experiment. Having shown that in settings where eye gaze traces are not available, substantial insight into topical relevance of terms can be gained from mouse cursor movements, the next section of this paper will demonstrate the use of this source of evidence for the task of query suggestion.

Table 8: Per-term cursor hover frequency and duration show a general upwards trend as the semantic proximity to query terms increases.

	hover duration	hover likelihood
$LCH < 0.25$	116ms	0.041
$0.25 \leq LCH < 0.5$	121ms	0.041
$0.5 \leq LCH < 0.75$	119ms	0.039
$0.75 \leq LCH < 1.0$	123ms	0.043
$1.0 \leq LCH < 1.25$	127ms	0.047
$1.25 \leq LCH < 1.5$	126ms	0.048
$1.5 \leq LCH < 1.75$	129ms	0.050
$1.75 \leq LCH < 2.0$	**133ms***	**0.053***
$LCH \geq 2.0$	**135ms***	**0.055***

6. MODELLING QUERY TERM USAGE

In the past, various successful applications integrated attention and interest information in the form of document display times [25, 23], clickthrough features [9, 45], hitting time [39], or the contents of personal data collections [10]. Most notably, Buscher et al. [8] use fixations on different document parts to reorder query expansion candidates. Their work proposed 4 different interest metrics, the best-performing one of which we will include as a baseline for comparison to our own method. Several advances in eye-tracking technology allow us to infer even more fine-grained signals than those studied by previous work. Concretely, **(1)** we map fixation durations and frequencies to individual terms while Buscher et al. rely on paragraph-level information. **(2)** In the previous sections, we saw that many reformulations are inspired by fixated terms, but often use related terms rather than the literal fixation term. In order to account for this effect, we include a model of semantic relatedness between candidate terms and fixation terms. Essentially, this broadens the coverage of our method and accounts for reformulations that include previously unseen terms in a fashion similar to language model smoothing.

6.1 Methodology

Buscher et al. [8] report best query expansion performance for their *Gaze-Length-Filter* (GLF). The method expands the well-known tf-idf formula by a user interest model based on the number of fixations on text segments shorter than 230 characters that contain a word w ($SA(w)$) and the frequency of longer text segments $LA(w)$ containing that same term. Their approach modifies the standard tf-idf formula in such a way, that only the frequency of w in those segments of the document that were gazed at (c_A) is considered. We include

Table 9: Query suggestion performance.

Method	MRR	σ_{RR}
API Output	0.76	0.28
Gaze-Length-Filter (GLF)	0.79*	0.26
Term-Attention-Model (TAM)	0.80*	0.24
Term-Attention-Model + Relatedness (TAM-R)	0.86$^{\blacktriangledown}$*	0.29

their model as a performance baseline. Please note that, in the following, we speak about words w rather than the previously discussed tokens t. While tokens are the atomic unit that receives measurable user attention, words represent general concepts. As a consequence, we add up the cumulative attention measured for each occurrence (token) t_w of a word w in order to estimate the word's relevancy to the user's information need.

$$GLF(w) = tf(w, c_A) \times idf(w, C) \times \frac{LA(w)}{LA(w) + SA(w)} \quad (4)$$

Additionally, we propose two novel, term-level attention models. $TAM_\lambda(w)$ combines $F(w)$, the frequency of attention to term w with the cumulative length for which the attention lasted $D(w) = \sum dur(t_w)$. The mixture parameter λ balances the relative contribution of the two terms, biasing the score towards attention frequency, as λ increases.

$$TAM_\lambda(w) = \lambda F(w) + (1 - \lambda) D(w) \quad (5)$$

At this point, neither the GLF baseline nor the TAM score can account for the addition of previously unseen terms w. In Sections 4 and 5, however, we observed significant shares of added query terms to not have been explicitly present or fixated before the reformulation. To remedy this, $TAM\text{-}R_\Lambda(w)$ expands TAM by a semantic similarity metric $LCH(w)$ between candidate term w and the set F of all terms that previously received user attention (fixation or cursor hover, respectively). In this way, we can ensure that the overall model score does not default to zero for unseen terms. The weight vector Λ defines the concrete relative contributions of semantic relatedness, attention frequency and attention duration.

$$TAM\text{-}R_\Lambda(w) = \lambda_l LCH(w) + \lambda_f F(w) + \lambda_d D(w) s \quad (6)$$

6.2 Experiments

Our experiments are based on another series of crowdsourcing tasks on our custom search engine interface. The experimental setup is identical to the one described earlier in Section 5. In this case, however, we included the top 7 query suggestions delivered by a commercial search engine API and measured the reciprocal rank (RR) of accepted (clicked) suggestions. We compare the original order of query suggestions with 3 alternative variants, each re-ranked by decreasing averaged scores in the three attention metrics (GLF, TAM, TAM-R) that were computed based on cursor hover information. Table 9 compares the query suggestion performance of the unmodified commercial API output with that of the various, previously introduced user attention-based models in terms of mean RR (MRR) across all accepted suggestions. Additionally, we inspect the stability of the methods in terms of their RR score variance.

For of all models that employ traces of user attention, we can observe consistently and significantly higher ranking performance with respect to the original API output. The mild performance gap, favouring TAM over the passage-based GLF could not be confirmed significant. As we, however, include the proximity-based smoothing functionality of TAM-R, query suggestion performance improves. Statistical significance of improvements was measured by means of a Wilcoxon signed rank test at $\alpha < 0.05$-level. Significant improvements over the baseline API output are denoted by an asterisk, while significant improvements over the GLF baseline are indicated by the $^{\blacktriangledown}$ symbol. Manual analysis of the re-orderings introduced by the various models confirms that GLF and TAM incorporate knowledge about literal term acquisitions from previously consumed material. TAM-R, indeed, accounts for the addition of terms that relate to the same topic that the user is interested in, but that did not directly occur in the explored document segments. The best-performing combination of mixture weights ($\lambda_l = 0.3$, $\lambda_f = 0.5$, $\lambda_d = 0.2$) was determined by means of a greedy parameter sweep in the range $[0, 1]$ with step size 0.1, ensuring $\sum \lambda_i = 1$ at all times.

6.3 Qualitative Performance Analysis

Besides the mere quantitative performance overview, we manually inspected those cases in which the various systems performed especially well (or badly) in order to give qualitative insight into the respective strengths and weaknesses of the presented methods. As reflected by the solid baseline performance, the raw API output in many cases returns the correct suggestion candidate on the highest ranks. Exceptions to this rule were broad queries that cover many potential aspects of a topic. As one out of several examples that we encountered, take Topic 38, that deals with psychological and emotional consequences of obesity. While our test subjects were mainly interested in mental problems that accompany this medical condition, top ranking query suggestions were concerned with vascular conditions as well as damage to the joints caused by dramatically increased weight.

All three user attention based models successfully placed such off-focus suggestions at the bottom of the candidate list, thereby increasing ranking performance. Both GLF and TAM struggled with previously unseen terms. Users that only briefly explored the available information before reformulating their query saw worst performance issues since their sparse observation vectors were not indicative towards the correct choice of suggestion.

TAM-R, finally, showed significant improvements in such situations. The use of the semantic proximity component served to assign probability mass to previously unseen terms. In practice, this resulted in promoting suggestion candidates that named concrete anabolic steroid substances such as the legally sold Prostanozol rather than the overall class. The remaining performance gap can mostly be accounted to pairwise candidate swaps between the top ranks of our list and the clicked suggestion. In many cases, this happened for near-duplicate candidates such as "*steroid consequences*" and "*steroid substance consequences*" in which the added term "*substance*" does not significantly modify the query semantics.

7. CONCLUSION

In this paper, we studied query reformulation by means of an eye-gaze tracking system. Inspired by topics drawn from the TREC QA track, we conduct a series of lab-based user studies. Tracking user attention at the term level, a finer granularity than was previously used in the IR literature, we make a range of interesting observations: **(1)** A significant share of newly added query terms were previously present on SERPs and visited pages in the same session. Previous work on the log-based recognition of query term acquisition [15] overestimated this effect. With the help of eye tracking hardware, we were able to gain a more realistic impression of how many such term occurrences were indeed seen by the user. **(2)** We find that literal query term acquisition is often indicated by significantly higher-than-average amounts of prior user attention in the form of frequency and duration of fixations to the prospective query terms. **(3)** Often, query expansion does not literally re-use previously encountered terms but highly related ones, instead. In a series of experiments we highlighted the importance of semantic proximity between query expansion terms and the center of user attention. **(4)** To ensure the broad applicability of our results, we replicated our lab-based eye-tracking experiments in a distributed fashion at much larger scale by measuring mouse cursor movements instead of eye gaze fixations. We note that our high-level findings generalize well between the two signals and the conclusions drawn from the lab-based study are confirmed by mouse cursor traces. **(5)** Finally, we demonstrate the usefulness of our findings for established IR tasks by comparing a passage-level attention model proposed by previous work to two variants of our term-level attention model, finding that term-level models including information about semantic proximity between candidate terms and user interest can deliver significantly better ranking performance of query suggestions than an industrial baseline.

The insights presented in this paper inspire many interesting directions for future work. First of all, the significant performance gains achieved by incorporating estimates of user attention into the query reformulation process motivate an evaluation of other related tasks. User attention models empowered by eye-gaze or cursor movement signals hold potential gains for ranking results to subsequent queries in a session, diversifying result sets, estimating domain expertise or personalized textual complexity. In this work, we used a very short-lived attention model based exclusively on the contents of the current search session. This was mainly due to limited availability of resources of users. Assuming an industrial setting, long-term attention models that include the searcher's general interest in addition to the current session context can be expected to become powerful tools for a wide number of inference tasks. In this way, one could estimate a general user vocabulary model, that describes the searcher's active and passive language use in more than just term frequencies. Such a model could for example describe the ease with which a user generates and consumes a given term, the speed at which they expand their vocabulary of new domains, or gradual shifts in interest. Wide-coverage models like that are especially interesting in query-free environments in which the system pro-actively pushes information about standing queries or upcoming events to the user.

In this work, we showed how models of knowledge acquisition in terms of previously unknown terms can benefit IR tasks. There are many other types of knowledge acquisition, e.g., factual or procedural knowledge, that can greatly benefit retrieval performance. Gaining an understanding of these learning processes holds significant potential for delivering smarter, more user-aware retrieval facilities. Finally, while the focus of this work lies on fixations, there are multiple other signals that can be captured by means of eye-tracking hardware. In our experiments, we additionally measured pupil dilation and saccade patterns. These signals turned out to be rather noisy and inconclusive when broken down on term level. For reasons of space, we omitted the respective results from the paper. In the future, we plan to conduct a dedicated investigation of these adjunct signals.

8. REFERENCES

[1] M. Ageev, D. Lagun, and E. Agichtein. Improving search result summaries by using searcher behavior data. In *SIGIR 2013*.

[2] A. Ajanki, D. Hardoon, S. Kaski, K. Puolamäki, and J. Shawe-Taylor. Can eyes reveal interest? implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction*, 2009.

[3] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*. Springer, 2002.

[4] N. Belkin, R. Oddy, and H. Brooks. Ask for information retrieval: Part i. background and theory. *Journal of documentation*, 1982.

[5] D. Beymer and D. Russell. Webgazeanalyzer: a system for capturing and analyzing web reading behavior using eye gaze. In *CHI 2005*. ACM.

[6] P. Brooks, K. Phang, R. Bradley, D. Oard, R. White, and F. Guimbretire. Measuring the utility of gaze detection for task modeling: A preliminary study. In *Workshop on Intelligent Interfaces for Intelligent Analysis*, 2006.

[7] G. Buscher, A. Dengel, and L. van Elst. Eye movements as implicit relevance feedback. In *CHI 2008*. ACM.

[8] G. Buscher, A. Dengel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. In *SIGIR 2008*. ACM.

[9] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In *KDD 2008*. ACM.

[10] P. Chirita, C. Firan, and W. Nejdl. Personalized query expansion for the web. In *SIGIR 2007*. ACM.

[11] K. Collins-Thompson, P. Bennett, R. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *CIKM 2011*. ACM.

[12] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *CHI 2007*. ACM.

[13] E. Efthimiadis. Query expansion. *Annual review of information science and technology*, 1996.

[14] C. Eickhoff, P. Serdyukov, and A. P. De Vries. A combined topical/non-topical approach to identifying web sites for children. In *WSDM 2011*.

[15] C. Eickhoff, J. Teevan, R. White, and S. Dumais. Lessons from the journey: A query log analysis of within-session learning. In *WSDM 2014*. ACM.

[16] W. Gao, C. Niu, J. Nie, M. Zhou, J. Hu, K. Wong, and H. Hon. Cross-lingual query suggestion using query logs of different languages. In *SIGIR 2007*. ACM.

[17] L. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR 2004*. ACM.

[18] Q. Guo and E. Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *WWW 2012*. ACM.

[19] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI 2010*. ACM.

[20] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 1998.

[21] J. Huang, R. White, G. Buscher, and K. Wang. Improving searcher models using mouse cursor activity. In *SIGIR 2012*. ACM.

[22] J. Huang, R. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *CHI 2011*. ACM.

[23] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR 2005*. ACM.

[24] M. Just and P. Carpenter. A theory of reading: From eye fixations to comprehension. *Psychological review*, 1980.

[25] D. Kelly and N. Belkin. Display time as implicit feedback: understanding task effects. In *SIGIR 2004*. ACM.

[26] D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. In *JCDL 2002*. ACM.

[27] D. Kelly, K. Gyllstrom, and E. Bailey. A comparison of query and term suggestion features for interactive searching. In *SIGIR 2009*. ACM.

[28] D. Kelly and J. Lin. Overview of the trec 2006 ciqa task. In *SIGIR Forum 2007*. ACM.

[29] J. Kim, K. Collins-Thompson, P. Bennett, and S. Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *WSDM 2012*. ACM.

[30] K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise. Towards inferring language expertise using eye tracking. In *CHI 2013*.

[31] K. Kunze, Y. Utsumi, Y. Shiga, K. Kise, and A. Bulling. I know what you are reading: recognition of document types using mobile eye tracking. In *Proceedings of the 17th annual international symposium on wearable computers*. ACM, 2013.

[32] V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 2012.

[33] T. Lau and E. Horvitz. Patterns of search: Analyzing and modeling web query refinement. *Courses and lectures-International Centre for Mechanical Sciences*, 1999.

[34] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 1998.

[35] J. Liu, N. Belkin, X. Zhang, and X. Yuan. Examining users' knowledge change in the task completion process. *IPM 2012*.

[36] J. Liu, M. J. Cole, C. Liu, R. Bierig, J. Gwizdka, N. J. Belkin, J. Zhang, and X. Zhang. Search behaviors in different task types. In *JCDL 2010*. ACM.

[37] T. Loboda, P. Brusilovsky, and J. Brunstein. Inferring word relevance from eye-movements of readers. In *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, 2011.

[38] H. Ma, H. Yang, I. King, and M. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In *CIKM 2008*. ACM.

[39] Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time. In *CIKM 2008*. ACM.

[40] M. Porter. Snowball: A language for stemming algorithms, 2001.

[41] K. Puolamäki, A. Ajanki, and S. Kaski. Learning to learn implicit queries from gaze patterns. In *ICML 2008*. ACM.

[42] K. Puolamäki, J. Salojärvi, E. Savia, J. Simola, and S. Kaski. Combining eye movements and collaborative filtering for proactive information retrieval. In *SIGIR 2005*. ACM.

[43] K Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 1998.

[44] J. Salojärvi, I. Kojo, J. Simola, and S. Kaski. Can relevance be inferred from eye movements in information retrieval. In *WSOM 2003*.

[45] Y. Song and L. He. Optimal rare query suggestion with implicit user feedback. In *WWW 2010*. ACM.

[46] A. Spink and T. Saracevic. Interaction in information retrieval: selection and effectiveness of search terms. *JASIS*, 1997.

[47] Amanda Spink. Term relevance feedback and query expansion: relation to design. In *SIGIR 1994*. Springer.

[48] J. Teevan, S. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR 2005*. ACM.

[49] R. White, S. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *WSDM 2009*. ACM.

[50] B. Wildemuth. The effects of domain knowledge on search tactic formulation. *JASIST 2004*.

[51] R. Williams and R. Morris. Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, 2004.

[52] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *ACL 1994*.

[53] X. Zhang, M. Cole, and N. Belkin. Predicting users' domain knowledge from search behaviors. In *SIGIR 2011*. ACM.