What Eyes Can Tell About the Use of Relevance Criteria During Predictive Relevance Judgment?

Panos Balatsoukas
Department of Computer and Information Sciences
University of Strathclyde, Glasgow, UK
+44 141 548 3092
panos@cis.strath.ac.uk

Ian Ruthven
Department of Computer and Information Sciences
University of Strathclyde, Glasgow, UK
+44 141 548 3098
ian.ruthven@cis.strath.ac.uk

ABSTRACT
This paper reports on the preliminary findings of a user study that explored how searchers fixate on information associated with different relevance criteria during the process of predictive relevance judgment. In order to address this objective a user study was conducted that involved the completion of questionnaires, use of eye tracking technology, talk aloud protocols and post-search interviews. As opposed to previous studies, the present research asked participants to search for real information needs that represented different search contexts (e.g. from searches about personal interest to academic related searches). This permitted the identification of several relevance criteria that naturally occur across different search contexts and the emergence of some fixation patterns, not observed before, associated to the use of these criteria. The paper concludes with a discussion of the impact and implication of this study in the wider context of relevance judgment and information seeking in context research.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval – Search Process.

General Terms
Measurement, Experimentation, Human Factors

Keywords
Human Relevance Judgment, Relevance Criteria, Eye tracking, User Studies.

1. INTRODUCTION
To date, several studies have investigated the use of relevance criteria by information users on the web e.g. [1, 2, 7]. These studies have shown that when participants search for information on the web they tend to make decision about its relevance based on several criteria, such as Topicality, Recency, Background knowledge, Scope and Utility. However, there is little known about what people actually see on the interface when they decide to evaluate the information sought based on one or more relevance criteria. Although eye tracking technology has provided the opportunity for researchers to examine what actually people look at when searching for information on the web [e.g. 3 gives a review], there are no studies exploring the relationship between relevance criteria use and human eye movements (e.g. number of fixations, fixation length and scan-paths). The purpose of this paper is to present the preliminary findings of a research that aims to address this gap by exploring the amount of visual attention spent by people fixating on information related to various relevance criteria during web searching. It is anticipated that the association of eye movement data (such as the number and length of fixations) with relevance criteria could enhance our understanding of how people make decisions on the web and improve the design of information and systems, such as web-pages, search result interfaces and relevance feedback mechanisms. The remainder of this paper presents the methodology and preliminary findings of this research (sections 2 and 3) as well as some discussion and conclusions (section 4).

IIiX 2010, August 18–21, 2010, New Brunswick, New Jersey, USA.

2. METHODOLOGY
The results reported in this paper came from the preliminary analysis of the data collected from 22 searches for information made by seven students, both undergraduates and postgraduates who were studying different disciplines at the University of Strathclyde. Specifically, there were 4 males and 3 females, all aged between 17 and 25 years old. Because the purpose of this study was to examine the association of eye fixations and relevance criteria use in different contexts, all participants searched for a real information need. No restrictions were imposed to them as far as concerns the type of search engines or web-pages they could use in order to find relevant information. Participants' information needs ranged from topics of personal interest (such as planning a trip to the Highlands) to academic-related topics (such as finding information about iPhone programming codes). The remainder of this section presents the research design implementation in more depth.

Participants were recruited through e-mails and announcements in university notice-boards. Students who expressed an interest in participating in the study were asked to complete a screening questionnaire in advance. The screening questionnaire ensured that participants had a good level of experience in web searching and did not suffer from any form of visual impairment. After the screening process, selected participants were contacted through emails and meetings were arranged in a laboratory specifically

designed for the purpose of this research. During the user study participants were asked to fill in a consent form and a background questionnaire. The purpose of the background questionnaire was to provide an in-depth knowledge about the context of participants' information need (such as the topic of their search, their interest in the particular topic, their motivation and confidence in judging relevance about that topic as well as the stage of participants' information seeking process). After the completion of the background questionnaire, participants were asked to search for information on the web about their information need as they would normally do. Each participant searched the web for 25 minutes. This time limit gave participants enough time to perform successive searches for information about their information need. Also, during the search process participants were asked to talk aloud, i.e. verbalise any thoughts or feelings that naturally came to their mind while searching for information [4]. Participants' eye movements were captured through the use of an eye tracking device (Tobii T60). The device consisted of a 17inch screen with the eye tracker embedded in it and permitted a 60Hz sampling rate, 0.5 degrees gaze point accuracy and free head motion.

After searching the web, participants were invited to participate in a semi-structured interview. During the interview participants watched a video of their searching behaviour (using the Tobii studio software tool) and responded to a fixed set of questions for each *predictive* and *evaluative* judgment [5] they made while searching the web. In the case of predictive judgments, participants were asked about the reasons why they decided to click on a specific link, what prevented them from selecting another option-link, what kind of information they expected to find by clicking on a selected link and finally, whether they expected to find very relevant, partially relevant, or not relevant information. Also, participants were asked to explain what words or phrases helped them reach a decision about the relevance of the information sought. Although the study was focused on both predictive and evaluative relevance judgments, the results reported in the next section are limited to the predictive relevance judgment process and specifically participants' evaluation of the results presented in the search result interface of the Google search engine.

Following a modified version of a coding scheme used by [6], the analysis of the interview transcripts and the talk aloud protocol gave rise to a set of relevance criteria used by participants (see section 3.1). These criteria were associated with specific words or phrases fixated by the participants and the fixation data collected from the eye tracker. More specifically the data collected for each relevance criterion were: the *total number of fixations*, the total *fixation length*, the *average fixation length* (that measures an average length per fixation), and the *average criterion fixation length* (as opposed to the average fixation length which is based on the division between the total fixation length and the total number of fixations, the average criterion fixation length is a more compound mean that divides the total length of fixations made for a specific relevance criterion by the total number of individual occurrences of this criterion, for example, one occurrence of a relevance criterion may consist of several fixations on, or re-visits to, a specific piece of information which is associated with that specific criterion).

3. RESULTS

3.1 Relevance Criteria and Eye Fixation Data

The data analysis showed that participants used a total of 12 relevance criteria in order to judge the relevance of the surrogates presented in the search result interface of the Google search engine. These were: *Topicality*; *Quality* (such as the reliability and quality of the contents, or the reputation of the author and the resource); *Recency*; *Format* (for example, whether the resource was available in a PDF or HTML format); *Tangibility* (covering relevance judgments about the utility of the information presented in the resource, or the amount and type of the data presented, such as the use of tables, use of raw data and figures); *Scope/Specificity/Depth* (this criterion covers judgments related to the scope, depth, completeness or level of specificity of the resource); *Resource type* (for example, whether it is an academic paper, an online tutorial or a personal website); *Affectiveness* (this criterion is related to judgments influenced by emotions evoked by the resource, such as interest or disappointment); *User background* (judgments influenced by individual-subjective factors, such as the level of familiarity with a resource, the level of knowledge possessed about the topic, or the perceived difficulty of the contents); *Document characteristics* (this criterion involves judgments related to the language or the version of the resource); *Serendipity* (this criterion represents a user's expectation to find accidentally relevant information in a distal web document which had been judged as not relevant during the predictive relevance judgment process); *Ranking* (this criterion represents a lack of analytical judgments, i.e the lack of use of relevance criteria such as those presented above. In this case, the decision to select, or reject, a specific surrogate is biased by its position in the ranked list. For example, whether the surrogate appears at the top or bottom of the page). Table 1 presents the means for all four measures (see section 2) across the 12 relevance criteria.

As it is shown in Table 1, participants made more and longer fixations (total number of fixations and total fixation length) on information that triggered the criterion of *Scope/Specificity*. This criterion was followed by the criterion of *Topicality* which scored marginally lower. Although participants made less fixations in the case of the criterion of *Resource type* (mean number = 0.5) than the criteria of *Topicality* and *Scope/Specificity* (6.1 and 6.4 fixations respectively), the former criterion was characterized by longer fixations at both the individual fixation length (average fixation length) and criterion level (average criterion fixation). This finding shows that whenever information in the search results triggered the criterion of *Resource type*, participants tended to pay attention for a longer period of time on information associated with this criterion than on information related to other types of relevance criteria. Similar was the data for the criteria of *Quality* and *Ranking*, as well as for the criterion of *Affectiveness* which scored higher than all the other criteria at the average criterion fixation level. In the case of the criterion of *Affectiveness* a big difference between the average fixation length (0.3 msecs) and average criterion fixation (1.2 msecs) can be observed. This difference, which is more evident for this criterion than in the case of the other criteria, shows that participants made more, but short, fixations whenever the criterion of *Affectiveness* occured.

Table 1. Fixation data per relevance criterion (mean values)*

Criteria	Fixation Number	Total Fix. Length	Average Fix. Length	Average Criterion Fix.Length
Affectiveness	0.8	0.2	0.3	1.2
User background	2.4	0.9	0.4	0.8
Document characteristic	0.4	0.1	0.2	0.5
Format	0.2	0.1	0.3	0.4
Quality	0.5	0.5	1.0	0.9
Ranking	0.1	0.1	1.3	1.4
Recency	0.7	0.3	0.4	0.6
Resource type	0.5	0.4	0.8	1.1
Scope	6.4	2.0	0.4	1.1
Serendipity	0.2	0.1	0.2	0.3
Tangibility	2.0	0.7	0.4	0.6
Topicality	6.1	2.1	0.3	0.9

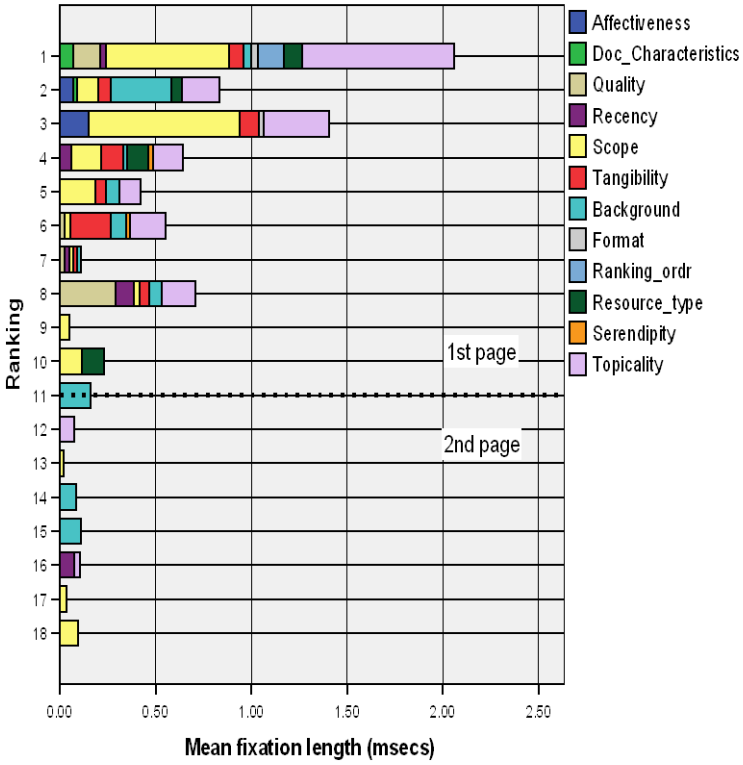
*Length values are reported in secs.

3.2 Relevance Criteria and Ranking Order

Participants used most relevance criteria for surrogates presented at the first page of the search results (Figure 1). However, the number or type of relevance criteria used was not equally distributed among the surrogates. In particular, participants used a total of 10 relevance criteria for the surrogates ranked first; seven criteria for surrogates ranked in positions two and four; six relevance criteria for surrogates in positions six and eight; five criteria for surrogates in ranked order three and seven; and finally, four or less criteria for surrogates ranked in positions five, nine and ten. In the case of the second search result page, participants tended to make judgments using only one or two relevance criteria. In the case of the second result page, the number of criteria used per surrogate was rather homogeneous. Figure 1 shows the distribution of the average fixation length per relevance criterion across 18 ranked surrogates in the Google search result interface.

As it is shown in Figure 1 participants tended to spent more time fixating on the Top-3 ranked surrogates. Also, there were differences in the time spent fixating on the various relevance criteria between ranked positions. Specifically, it appears that as the participants moved from the top to the bottom of the search result list, the time spent fixating on information related to the criterion of Topicality decreased. At the same time, as the interest in Topicality decreased, participants tended to spent more time fixating on other types of criteria in order to judge the relevance of the surrogates, such as the criterion of Scope/Specificity (which dominates in the case of surrogates ranked in positions 3, 4, 5, 9 and 10), or the criteria of tangibility and quality (these criteria dominated in the case of surrogates ranked in positions 6 and 8 respectively). Finally, it appears that participants viewed less frequently the surrogates presented in the second page of the search results. However, when this happened they tended to fixate on information related to the criteria of Scope/Specificity and User Background in order to make decisions about the relevance of the retrieved surrogates. Similar were the findings in the case of the mean number of fixations per relevance criterion across the ranked surrogates.

Figure 1. Averaged Total Fix. Length across search results.



3.3 Relevance Criteria and Surrogate Components

Data was also analyzed at the surrogate level. As it is shown in Table 2 participants tended to use different types of relevance criteria when fixating on the Title, Summary or the URL component of a surrogate. For example, participants made more fixations and spent more time fixating on information cues that triggered the criterion of Topicality and Scope/Specificity in the Title rather than the Summary or the URL. Also, information cues related to the criteria of User Background (such as familiarity with the resource) and Quality were fixated more often and for a longer period of time in the case of the URL element of the surrogate than the Title or the Summary. Likely, this happened because participants made judgments about the quality or their level of familiarity with the distal web document just by viewing the domain name of the URL (e.g. www.wikipedia.com, or http://www.acm.org/). Furthermore, information cues that triggered the criteria of Affectiveness, Recency and Tangibility were fixated more often and for a longer period of time at the Summary level of the surrogate rather than the Title and the URL. This behaviour can be explained by the fact that Google's automatically generated summary usually provides more information than the title or the URL, such as dates (like in the case of the criterion of Recency), information about the utlility, type or amount of data (like in the case of the criterion of Tangibility), or unexpected points of view or information related

to the topic under investigation (like in the case of the criterion of Affectiveness).

Table 2. Fixation data across surrogate components*

	Title		Summary		URL	
Criteria	**N(f)**	**Length (secs)**	**N(f)**	**Length (secs)**	**N(f)**	**Length (secs)**
Affective-ness	0.3	0.1	0.5	0.1	0.0	0.0
Document characteri-stics	0.4	0.1	0.0	0.0	0.0	0.0
Quality	0.1	0.0	0.0	0.0	0.4	0.5
Recency	0.0	0.0	0.7	0.3	0.0	0.0
Scope	4.7	1.7	1.4	0.5	0.3	0.1
User Background	0.4	0.2	0.7	0.3	1.3	0.5
Format	0.2	0.1	0.0	0.0	0.0	0.0
Ranking	0.0	0.0	0.1	0.1	0.0	0.0
Resource type	0.1	0.1	0.2	0.2	0.2	0.1
Serendipity	0.1	0.0	0.0	0.0	0.1	0.0
Topicality	3.5	1.1	2.1	0.6	0.4	0.2
Tangibility	0.7	0.3	1.2	0.3	0.1	0.0

* N(f) = mean Number of Fixations; Length(secs) = mean Total Fixation Length.

4. DISCUSSION - CONCLUSIONS

The purpose of this paper was to present some preliminary findings that could shed light on the process of predictive relevance judgment and especially on the problem of relevance criteria use. To date, studies on relevance criteria were limited to the use of retrospective or concurrent think aloud protocols and interviews that provided data about the frequency with which specific criteria occurred during the search session [e.g. 2; 6]. However, the present study goes a step further by associating the use of relevance criteria to behavioural data (such as the number of fixations and fixation length) and interface components (such as the use of relevance criteria across ranked results and within the various surrogate components). Also, unlike previous eye tracking studies which made use of eye tracking as the only data collection technique [3] and were focused on the performance of controlled and simulated search tasks [3, 7], the present research validates the suitability of using eye tracking in a multi-method research design (data was collected from talk aloud protocols, retrospective interviews and eye tracking) that involved real tasks for studying predictive relevance judgment behaviour in different search contexts.

Since the results reported on this paper are preliminary, any generalisations should be made with caution. However, some patterns seem to appear as far as concerns the use of certain relevance criteria and users' eye movements. For example, participants in the study tended to spend more time fixating on information associated with the criteria of Topicality and Scope/Specificity. Also, participants applied analytic judgment techniques even for surrogates ranked below the Top-2 results. Analytical judgment techniques involve the combination of one or more relevance criteria in order for a user to reach a decision about the relevance of a retrieved surrogate. Finally, different surrogate components (Title, Summary and URL) seem to be associated with fixations that trigger different types of relevance criteria (e.g. see Table 2). These findings were not shown in previous studies that examined users' interaction with the search result interface of the Google search engine [e.g. 3], thus providing new knowledge about human relevance judgment behaviour in the context of information seeking research. Apart from its theoretical implications, this study could have implications also in the design of relevance feedback mechanisms and personalised search result interfaces for contextual information seeking. For example, work is in progress to examine in more depth the effects of contextual factors, such as the level of interest in the topic, type of search task and level of confidence in judging relevance on the interaction between relevance criteria use and users' eye movements. It is anticipated that the emergence of different fixation patterns could be captured for the development of implicit feedback mechanisms and the dynamic display of search results in context.

5. ACKNOWLEDGMENTS

This research is funded by the AHRC.

6. REFERENCES

[1] Crystal, A., and Greenberg, J. 2006. Relevance criteria identified by health information users during web searches. *Journal of the American Society for Information Science and Technology* 57, 10, 1368-2882.

[2] Tombros, A., Ruthven, I, and Jose, J. 2005. How users assess web pages for information seeking. *Journal of the American society for Information Science and Technology* 56, 4, 327-344.

[3] Loringo, L., Haridasan, M., Brynjarsdottir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F., Pan, B. 2008. Eye tracking and online search: lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, 59, 7, 1041-1052.

[4] Ericsson, A., and Simon, H. 1993. *Protocol analysis: verbal reports as data*. MIT press, Cambridge, Mass.

[5] Rieh, S.Y. 2002. Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology* 53, 2, 145-161.

[6] Barry, C. 1998. Document representations and clues to document relevance. *Journal of the American Society for Information Science and Technology*, 49, 14, 1293-1303.

[7] Papaeconomou, C., Zijlema, A. and Ingwersen, P. 2008. Searchers' relevance judgments and criteria in evaluating web pages in a learning style perspective. IIiX 2008. Information Interaction in Context 2008. 123-132.

What eyes can tell about the use of relevance criteria during predictive judgment?

Panos Balatsoukas and Ian Ruthven
Department of Computer and Information Sciences, University of Strathclyde, Glasgow

INTRODUCTION

To date, several studies have investigated the use of relevance criteria by information users on the web [e.g. 1, 2]. However, there is little known about what people actually see on the interface, and for how long, when they decide to evaluate the information sought based on one or more relevance criteria. The purpose of this research is to address this gap by exploring the amount of visual attention spent by people fixating on information related to various relevance criteria during their interaction with the search result interface of the Google search engine.

METHODOLOGY

The results reported in this paper came from the preliminary analysis of the data collected from 22 searches for information made by 7 students using Google search engine. The research design involved the completion of questionnaires, use of eye tracking technology, talk aloud protocols and post-search interviews. As opposed to previous studies, the present research asked participants to search for real information needs that represented different search contexts (e.g. from searches about personal interest to academic related searches). The analysis of the interview transcripts and the talk aloud protocol gave rise to a set of relevance criteria used by participants. These criteria were associated with specific words or phrases fixated by the participants, the location of the fixations (surrogate components and ranking position), participants' relevance judgments and the fixation data collected from the eye tracker (number and length of fixations) (Figure 1).

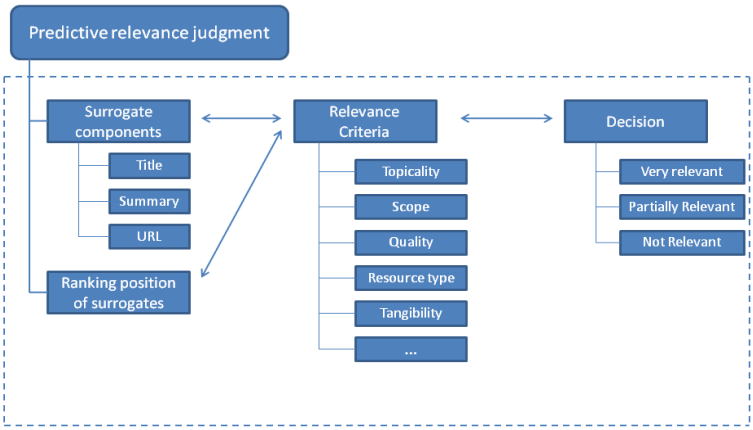


Fig.1. Context of the data analysis.

RESULTS

Relevance criteria and fixation data:

Participants made more and longer fixations (mean number of fixations=6.4 and mean fixation length=2.0secs) on information that triggered the criterion of *Scope/Specificity*. This criterion was followed by the criteria of *Topicality* and *User Background* which scored marginally lower.

Fig.2. Most criteria occurred at the top of the list

Relevance criteria and ranking order:

Participants used most relevance criteria for surrogates presented at the top of the first page of the search results (Figure 2). As the participants moved from the top to the bottom of the search result list, the time spent fixating on information related to the criterion of Topicality decreased. At the same time, as the interest in *Topicality* decreased, participants tended to spent more time fixating on other types of criteria in order to judge relevance, such as the criterion of *Scope/Specificity* (ranked positions 3, 4, 5, 9 and 10), or the criteria of *Tangibility* and *Quality* (ranked positions 6 and 8).

Relevance criteria and surrogate components:

Title: participants made more fixations (mean number = 4.1) and spent more time fixating (mean = 1.4secs) on information cues that triggered the criterion of *Topicality* and *Scope/Specificity*.

URL: The criteria of *User Background* (mean fixation number=1.3; mean length=0.5secs) and *Quality* (mean=0.4; mean= 0.5secs) were fixated more often and for a longer period of time.

Summary: Participants made more and longer fixations on information related to the criteria of *Affectiveness* (mean number of fixations=0.5; mean length=0.1secs), *Recency* (mean number=0.7; mean length=0.3secs) and *Tangibility* (mean number=0.7; mean length= 0.3secs).

FURTHER WORK

Work is in progress to examine in more depth the effects of contextual factors, such as the level of interest in the topic, type of search task and level of confidence in judging relevance on the interaction between relevance criteria use and users' eye movements. Fixation patterns could be captured for the development of implicit feedback mechanisms and the dynamic display of search results in context.

REFERENCES

[1] Crystal, A., and Greenberg, J. 2006. Relevance criteria identified by health information users during web searches. *Journal of the American Society for Information Science and Technology* 57, 10, 1368-2882.

[2] Tombros, A., Ruthven, I, and Jose, J. 2005. How users assess web pages for information seeking. *Journal of the American society for Information Science and Technology* 56, 4, 327-344.