The use of relevance criteria during predictive judgment: an eye tracking approach

Panos Balatsoukas
Department of Computer & Information Sciences
University of Strathclyde, Glasgow, UK
panos@cis.strath.ac.uk

Ian Ruthven
Department of Computer & Information Sciences
University of Strathclyde, Glasgow, UK
Ian.ruthven@cis.strath.ac.uk

ABSTRACT

This paper reports on the findings of a user study that explored how searchers tend to fixate on information associated with different relevance criteria in the web during the relevance judgment process. The user study 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. Although the study examined participants' eye movements for both predictive and evaluative relevance judgments, this paper is focused on the findings of the predictive relevance judgment process and specifically participants' evaluation of the results presented in the search result interface of the Google search engine. It is anticipated that the findings reported in this paper could shed light on the process of predictive relevance judgment and especially on the problem of relevance criteria use through the use of eye tracking.

Keywords

Human Relevance Judgment, Relevance Criteria, Eye Tracking, User Studies.

INTRODUCTION

The application of eye tracking technology has enabled researchers in information retrieval to gain a more in depth view of human relevance judgment behaviour in the web (Loringo et al., 2008) by providing behavioural data associated to the cognitive process of reasoning and decision making, such as the number of fixations, fixation length and scanpaths (Rayner, 2009). In this manner, several researchers have applied eye tracking techniques to examine human behavior in search engines (e.g. Loringo et al., 2008), web-pages (e.g. Papaeconomou et al., 2008) and corporate databases (e.g. Goldberg et al., 2002). In particular, in the context of *predictive relevance judgment* (Rieh, 2002) some empirical patterns of eye movements have emerged, such as the element of bias towards the top ranked results (Pan et al., 2007), the impact of task type on the number and length of fixations when evaluating search results (Loringo et al., 2006), or, the application of exhaustive and economic visual search strategies for the inspection of search result lists (Aula et al., 2005). *Predictive judgment* can be referred to any prediction made about a distal object in the web (for example, a web-page) based on information contained in a proximal object, such as a surrogate in a search result interface or a link in a web-page. Many researchers have advocated the importance of understanding predictive judgments in order to improve the design of search engines and pages on the web and optimize the time and the accuracy of decision making (Pirolli, 2007). For many years, an important cluster of this research effort has been focused on the identification and examination of the criteria users apply when judging relevance (Barry and Schamber 1998). Although various types of relevance criteria have been identified in the context of the web, such as *Topicality*, *Scope*, *Tangibility*, *Recency* or *Format* (e.g. Rieh, 2002; Tombros et al., 2005; Crystal and Greenberg, 2006), there is little known about their actual use during the predictive judgment process and especially the attentional and cognitive effort spent on information related to these criteria. For example, is there an association between relevance criteria and specific components of a search results interface that people tend to fixate? or, What is the relationship between ranking of search results and the time spent fixating on information related to specific relevance criteria?

The purpose of this study was to address such questions by exploring the relationship between relevance criteria use and human eye movements (e.g. number of fixations, fixation length and scanpaths) during the process of predictive relevance judgment. In order to address this objective a user study was conducted that involved the collection of eye movement data during the evaluation of the results presented in the search result interface of the Google search engine. It is anticipated that the association of eye movement data with relevance criteria could enhance

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

our understanding of how people make decisions on the web and inform the design of information and systems, such as web-pages, search result interfaces and relevance feedback mechanisms. The paper is structured as follows. The next section presents some key studies that made use of eye tracking for the examination of human relevance judgment behavior in the context of various types of search result interfaces. Then, the methodology and findings of this research are reported. Finally, the last two sections present some discussion and conclusions.

LITERATURE REVIEW

Contextual factors and eye movements

To date, eye tracking has been used for the investigation of the effects of various contextual factors on human visual searching behaviour during the relevance judgment process. For example, Loringo et al. (2006) examined the impact of task type and gender on users' interaction with search results. For the purpose of this study, a total of 23 participants were recruited and performed a set of 10 navigational and informational search tasks using Google. Data collected included the number of fixations, duration of fixations and scan-path patterns. The researchers found that more male than female participants tended to fixate on the results presented near the bottom of the result page. However, female participants fixated more often on the results displayed at the Top-3 positions of the search result interface. Furthermore, males tended to view more result pages and their scan-path pattern was more linear than females. Also, differences were observed in the case of the type of search task. For example, participants spent significantly more time viewing search results in the case of navigational rather than informational tasks. Besides task and gender, other researchers have shown that further contextual factors, such as learning style can have an impact on the number and length of fixations performed during the evaluation of search results (Hughes, et al., 2003).

Presentation of search results and eye movements

Researchers have also reported on the importance of several search result interface characteristics for understanding human visual search behavior during relevance judgment, such as ranking position, length of surrogates and clustering of search results. In particular, Cuttrell and Guan (2007) examined the effects of ranking position and surrogate length of the MSN search engine on the web searching behaviour of 22 participants. All participants used the MSN search engine in order to perform a set of 12 informational and navigational tasks. The researchers collected common eye tracking data, such as fixation duration and number of items fixated as well as the number of snippets viewed before and after users had clicked on a selected item from the result list. The findings of this study showed that there was a significant effect of the interaction between type of task and surrogate length on the number of results fixated and the duration of fixations. For example, participants, who were presented with longer surrogates, tended to look at more hits in the result list in the case of navigational tasks rather than informational tasks. Also, the results showed an effect of ranking on the number of the results viewed and the duration of eye fixations. For example, participants viewed significantly more results presented near the top rather than the bottom of the first page of the search results. Similar findings were reported by Granka et al (2004) who found that participants in their study made more and longer fixations for the results ranked higher in the search result interface and especially the Top-2 results. Pan et al (2007) also found that ranking position can bias users' visual search and relevance judgment behavior when inspecting lists of search results. They reached this conclusion after analyzing data collected from 12 participants who performed a set of 10 informational and navigational tasks using three different types of the Google search engine. In the *Normal* type results were displayed in their original ranking order. In the *Swapped* condition the position of the Top-2 results was reversed and finally, in the *Reversed* type the Top-2 results were swapped with the results in positions nine and ten (i.e. the last two surrogates at the bottom of the first result page). Data collected included the number and length of fixations and pupil dilation. The findings showed that fewer participants completed correctly the tasks in the Reversed condition compared to the Normal and Swapped result interfaces, thus showing that participants tended to trust the results ranked higher in the list.

In another study Rele and Duchowski (2005) investigated the hypothesis that tabular search result interfaces can provide users with more accurate and efficient scanning of metadata surrogates than list interfaces. The researchers observed the ocular behavior of 16 test participants while performing four navigational and informational across the two interfaces of a pilot search engine. Data collected included time, errors, eye movement transitions, number of fixations and mean fixation duration. The results did not reveal any significant differences between the two interfaces in terms of time, errors and mean fixation duration. However, differences were observed between the two types of interfaces in the number of fixations made on the URL component of the surrogates. In particular, participants tended to view more frequently the URLs presented in the list rather than the tabular interfaces. Also, more fixations occurred in the surrogate summaries for the navigational tasks rather than the information tasks. Finally, the tabular organization of search results had an effect on participants' scan-paths who evaluated the search results vertically within columns rather than horizontally (between columns).

Apart from textual search result interfaces, other researchers have focused their research on visual search engines. For example, Tseng and Holmes (2008) found significant effects of the density and number of search result pages on participants' eye movements in the context

of image search engines, while Hughes et al (2003) found that participants spent significantly more time fixating on textual rather than visual components of metadata surrogates in the case of a video search engine.

Visual search patterns in search result interfaces

Visual search patterns and strategies during the process of relevance judgment have been reported by Aula et al (2005) and Rodden et al (2008). Aula et al (2005) investigated users' strategies when interacting with search result interfaces of web search engines. They studied 28 participants who viewed 10 search result pages, each page corresponding to a predefined search task. In particular, the researchers examined users' visual search behaviour for the period between the presentation of the search result page and users' first action (for example, selecting a result from the list, refining a query or entering a new URL). Data collected included fixation duration, number of fixations and users' scanpaths. The analysis showed that participants employed economic and exhaustive visual search strategies when evaluating search results. Economic evaluators tended to view fewer surrogates in the results page than the exhaustive evaluators. Also, economic evaluators were more experienced computer users and spent less time fixating on the results than the exhaustive evaluators. Moreover, the economic evaluators did not tend to evaluate surrogates displayed below the one clicked, while exhaustive evaluators inspected results more thoroughly and were not focused on the top results or the relevant only. These findings were confirmed by Dumais et al (2010) who suggested a method of categorizing participants as either economic or exhaustive evaluators based on their eye movements. In another study, Rodden et al (2008) investigated co-ordination patterns between eye and mouse movements during the examination of search results. Data were collected from a sample of 32 participants who performed 16 fact-finding tasks using the Google search engine. The data analysis showed the presence of three types of patterns: 1. Following the eye vertically with the mouse, 2. Following the eye horizontally with the mouse, and 3. Using the mouse to mark a promising result.

Relevance criteria

The study of relevance criteria use is associated to a cognitive approach to relevance. According to this approach, people do not tend to judge relevance solely based on topicality but also based on cognitive, socio-cognitive, situational and affective factors. These factors or *dimensions of relevance* are associated to different relevance criteria which people use and cognitively process in order to judge the relevance of information on the web (Spink, 2002). Researchers have investigated the use of relevance criteria in the context of both web search engines and structured with metadata information retrieval systems. This type of research helped the understanding of the multi-dimensional and dynamic nature of relevance. For example, users do not only judge relevance based on the topic. Other criteria used for relevance judgment may include the currency, quality, authority and availability of the resource as well as users' background and characteristics (Barry and Schamber, 1998; Rieh, 2002; Tombros et al., 2005).

Although many researchers have examined the use of relevance criteria on the web and search result interfaces in particular, the literature review showed a lack of studies investigating the association between eye movements and the use of relevance criteria during predictive relevance judgment. In an interesting study, Papaeconomou et al (2008) used eye tracking to examine how users of different learning styles apply relevance criteria to evaluate the usefulness of web-pages. However, the researchers did not report any data associating eye movements with relevance criteria use. In addition, the study by Papaeconomou et al was not focused on the context of predictive relevance judgment and search result interfaces in particular.

METHODOLOGY

The results reported in this paper came from the analysis of the data collected from 56 searches for information made in Google by 17 university students, both undergraduates and postgraduates, who were studying different disciplines. Specifically, there were 9 males and 8 females, all aged between 17 and 32 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. 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). No restrictions were imposed to participants as far as concerns the type of search engines or web-pages they could use in order to find relevant information. However, it is worth mentioning that only the data collected from user interaction with the search result interface of Google are reported in this paper. Since participants used a wide range of systems and web pages to search for information, it was anticipated that the focus of the analysis on a specific type of interface (e.g. Google) would provide the opportunity for a more focused and in depth analysis of human behavior before moving the analysis to a larger scale (i.e. before bringing together data collected from participants' interaction with different systems and web-pages). In addition, the decision to focus the analysis on Google was made because this system was selected by all participants in the study.

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. The recruitment of experienced web searchers served the need of reducing between subject variability by focusing on a sample of participants familiar with the information searching

process. In particular, qualified participants used the web for more than 10 hours per week, had *a lot* or *a very great deal* of experience with web searching and they were *frequent* or *very frequent* users of search engines, library catalogues and bibliographic databases.

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). However, results about the contextual data and their effect on participants' eye movements and judgment behavior will be presented in follow up publications. 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. In particular, each participant performed an average of 3.2 searches. It is worth mentioning that each query was considered to be a differed search.

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 (Ericsson and Simon, 1993). In order to eliminate bias of post-hoc rationalisation or reconstruction effects, participants were not prompted by the researchers to provide any explanations, descriptions and reasoning about the tasks performed at an earlier stage. However, when participants paused for more than 40 seconds they were reminded to keep talking. 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 behavior (using the Tobii studio software tool) and responded to a fixed set of questions for each *predictive* and *evaluative* judgment (Rieh, 2002) 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 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. As well as to selected surrogates participants were asked about the reasons why they decided not to click on the other surrogates that they fixated on. Also, participants were asked to explain what words or phrases helped them reach a decision about the relevance of the information sought (for all fixated surrogates). Although the study was focused on both predictive and evaluative relevance judgments, this paper is focused on the findings of the predictive relevance judgment process and specifically participants' evaluation of the results presented in the search result interface of the Google search engine. Finally, it is worth mentioning that indicative actions of predictive judgment in the Google search engine were: 1. The selection of a specific surrogate from the list, or, 2. The inactivity after the inspection of the search results (i.e. no surrogate selection). In this latter case, participants were asked about the reasons why they decided not to click on a surrogate.

Following a modified version of a coding scheme used by (Crystal and Greenberg, 2006), the analysis of the interview transcripts and the talk aloud protocol gave rise to a set of relevance criteria used by participants (see Table 1). These criteria were associated with specific words or phrases fixated by the participants, the location of the fixations and the fixation data collected from the eye tracker (Figure.1).

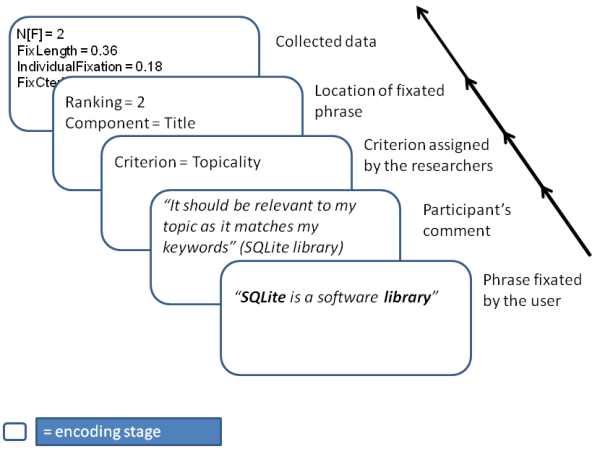

Figure. 1. Example of the coding and analysis process

More specifically the data collected for each relevance criterion were: the *number of fixations (N[f])*, the *fixation duration (FixLength)*, the *length of individual fixation* (that measures an average length per fixation), and the *average criterion fixation length (FixCriterion)* (as opposed to the average *length of individual fixation* 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).

RESULTS

Types of relevance criteria and 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 (Table 1). Table 2 presents the means for all four measures across the 12 relevance criteria.

Topicality	This criterion represents the topical relatedness of the material to users' queries.
Quality	The reliability and quality of the contents, or the reputation of the author and the resource.
Recency	How current, recent or up to date the material is.
Format	Format of the resource (e.g. PDF or HTML).
Tangibility	It covers the utility of the contents the resource, or the amount and type of the data presented (e.g. use of tables, raw data and figures.
Scope	This criterion covers judgments related to the scope, depth, completeness or level of specificity of the resource.
Resource type	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	The decision to select or reject a specific surrogate is biased by its position in the ranked list, e.g, whether it appears at the top or bottom of the page.

Table 1. Relevance criteria used by participants

As it is shown in Table 2, participants made more fixations (*N[f]*) and spent more time fixating (*FixLength*) on information related to the criterion of *Topicality*. Although the results showed that the criterion of *Scope* scored lower than *Topicality* for these two measures, it performed better than *Topicality* at the individual fixation (*length of individual fixation*) and criterion occurrence level (*FixCriterion*). This finding shows that the fixations made on information that is related to the criterion of *Topicality* were more, but shorter in average length, than the fixations made in the case of the criterion of *Scope*.

Criteria	N[f]	FixLength	Length of Individual fixation	FixCritetion Length
Affectiveness	0.9	0.3	0.3	0.8
Document Characteristics	0.2	0.1	0.3	0.5
Quality	0.9	0.4	0.4	0.8
Recency	0.5	0.2	0.4	0.7
Scope	3.1	1.6	0.5	1.1
Tangibility	1.3	0.2	0.1	0.3
User background	1.7	0.7	0.4	0.9
Format	0.1	0.0	0.4	0.4
Ranking	0.1	0.1	0.9	1.7
Resource type	0.3	0.2	0.6	0.9
Serendipity	0.1	0.0	0.3	0.3
Topicality	4.9	1.8	0.4	0.9

Table 2. Fixation data per relevance criterion (mean values, all *length* values in secs)

This behavior shows that information related to the latter criterion increased the cognitive effort spent by participants during the evaluation of the search results. Also, the large number of fixations for the criterion of *Topicality* can be attributed to the contents of the surrogates presented in the Google search engine, which can contain rich information about the topic of a retrieved document, like the information displayed on the Title or the short summary of the surrogate. It appears that this topical information was processed faster by participants who needed it in order to evaluate the extent to which a retrieved document matched their search topic. A similar observation can be made between other criteria, such as the criteria of *User Background* and *Resource type*. Participants made more fixations on information related to the criterion of *User background* rather than the criterion of *Resource type*. However, whenever participants attended on information related to the *Resource type* criterion, their fixations were longer, but fewer, than in the case of the criterion of *User Background* (Table 2).

Relevance criteria and ranking order

Participants used most relevance criteria for surrogates presented at the first page of the search results (Figure 2). 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 three, six, seven and eight, 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 used to make judgments using one

or two relevance criteria, the type of which was rather homogeneous.

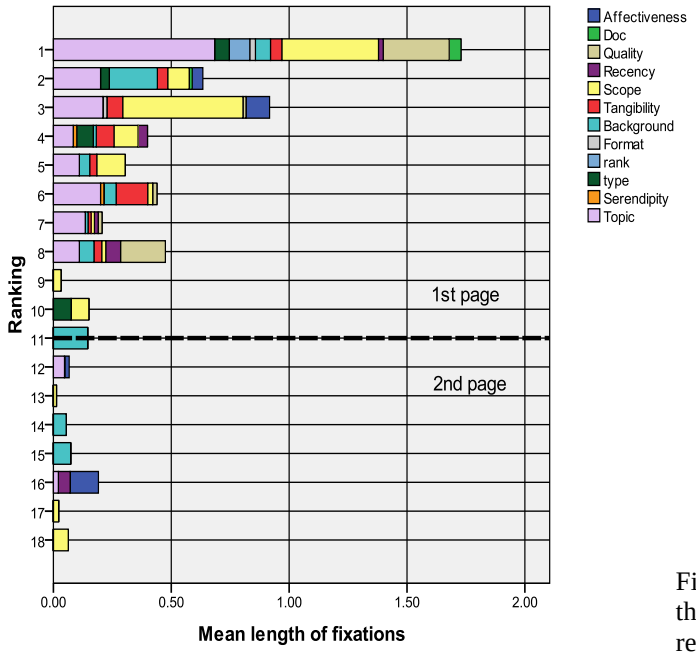

Figure 2. Mean fixation length (FixLength in secs) across search results.

Figure 2 shows the distribution of the average fixation length (*FixLength*) per relevance criterion across 18 ranked surrogates in the Google search result interface. Participants tended to spend more time fixating on the surrogates presented at the top of the search result list, and especially the Top-3 ranked surrogates. This was statistically confirmed by the results of the Pearson correlations which showed significantly negative relationships, at the 0.01 level, between the ranking of the results in the list and the means for fixation length (*FixLength: r = - .741, N = 18, p<0.01*) and number of fixations (*N[f]: r = -.731, N=18, p<0.01*) for all types of relevance criteria. Figure 3 displays this trend in a typical heat-map visualisation of users' eye movements.

Also, differences were observed in the time spent fixating (*FixLength*) 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 (Figure 2). 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* (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).

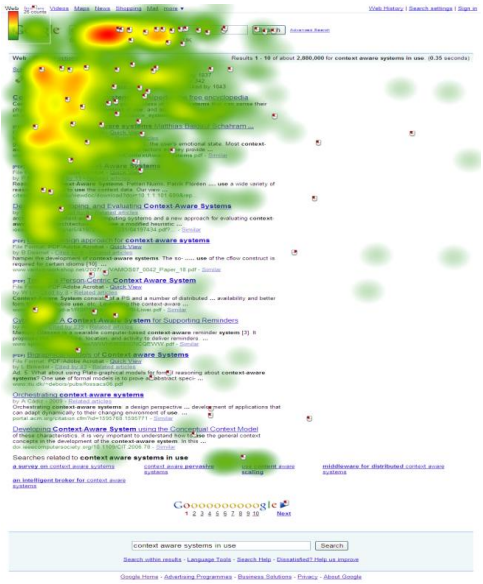

Figure 3. Heat map image of participants' fixations

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* and *User Background* in order to make decisions about the relevance of the retrieved surrogates. Similar were the findings in the case of the number of fixations (*N[f]*) per relevance criterion across the ranked surrogates as well as for the measures that dealt with the length of individual fixations and fixation criterion occurrence length (*FixCriterion*).

Relevance criteria and surrogate components

The findings showed differences in the number (*N[f]*) and length (*FixLength*) of fixations between the Title, Summary and URL components of the surrogates presented in the search result interface of the Google search engine. In particular, participants made a mean number of seven fixations and spent an average of 2.8 seconds fixating on the Title components, five fixations and 1.9 seconds on the Summary, and 1.7 fixations and 0.8 seconds fixation time on the URL. The results of the one-way within-subjects ANOVA tests showed that these differences were significant at the 0.01 level for the *N[f]*, and *FixLength* measures (*N[f]: p<0.0005*; *FixLength: p<0.0005*). However, there were no statistically significant differences in the case of the *length of individual fixations* and *FixCriterion* measures.

Also, differences were observed as far as concerns the use of relevance criteria across the three surrogate components. For example, participants made more fixations and spent more time fixating on information cues related to the criterion of *Topicality* and Scope in the Title rather than the Summary or the URL (Table 3). 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. 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 www.acm.org). Furthermore, information related to 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 behavior 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 but interesting points of view about the topic under investigation (like in the case of the criterion of *Affectiveness*). Although Table 3 displays only data about the number of fixations (*N[f]*) and fixation length (*FixLength*), these values had a similar distribution in the case of the other two measures (*individual fixation length* and *FixCriterion*).

	Title		**Summary**		**URL**	
Criteria	**N(f)**	**Fix Length secs**	**N(f)**	**Fix Length secs**	**N(f)**	**Fix Length secs**
Affectiveness	0.3	0.1	0.6	0.2	0.1	0.0
Document Characteristics	0.2	0.1	-	-	-	-
Quality	0.3	0.1	0.3	0.1	0.3	0.3
Recency	-	-	0.5	0.2	-	-
Scope	2.1	1.1	0.9	0.3	0.2	0.1
User Background	0.4	0.2	0.5	0.2	0.8	0.3
Format	0.1	0.0	-	-	-	-
Ranking	-	-	0.1	0.1	-	-
Resource type	0.1	0.1	0.1	0.1	0.1	0.1
Serendipity	0.1	0.0	-	-	0.1	0.0
Topicality	3.0	1.2	1.9	0.6	0.3	0.1
Tangibility	0.8	0.2	0.5	0.2	0.1	0.0

Table 3. Fixation data across surrogate components (mean values)

Relevance criteria and relevance judgments

Table 4 presents a breakdown of the eye movements (*N[f]* and *FixLength*) based on whether a participant decided to click or not to click on a surrogate. In the former case (where the participant decided to click) data are divided into three conditions: 1. The user clicked because expected to find very relevant information; 2. The user clicked because expected to find partially relevant information; and 3. The user decided to click but expected to find not relevant information (for example, a participant decided to click on a surrogate because it appeared from the abstract that it treated an interesting perspective of the topic under investigation. However, based on the transcript of the interview that followed the information searching session, this participant did not expect to find relevant information on the topic by clicking on the particular surrogate).

As it is shown in Table 4, *Topicality* was the dominant criterion for deciding to click on a surrogate which the user expected to lead to *very relevant* or *partially relevant* information (the Bonferroni table produced for the one-way within-subjects ANOVA test showed that the criterion of *Topicality* differed significantly from all the other criteria at the 0.05 level, with p. values ranging between $p<0.018$ and $p<0.0005$ for the *FixLength* measure, and between $p<0.046$ and $p<0.0005$ for the *N[f]* measure). In the case of surrogates characterized as *not relevant* (both selected and not selected) participants made more fixations and spent more time fixating on the criteria of *Scope*, *User background* and *Affectiveness* than the criterion of *Topicality*. In particular, in the case of the *not selected surrogates*, the Bonferroni tables showed that the difference between the criteria of Scope and Topicality was significant at the 0.05 level for both N[f] and FixLength measures (FixLength: $p<0.045$; F[N]: $p<0.03$).

	Selected Very relevant		**Selected Part.Relevant**		**Selected Not Relevant**		**Not Selected**	
Criteria	**N(f)**	**Fix Length secs**	**N(f)**	**Fix Length secs**	**N(f)**	**Fix Length secs**	**N(f)**	**Fix Length secs**
Affectiveness	0.1	0.0	-	-	0.3	0.1	0.5	0.1
Document characteristics	-	-	0.1	0.0	-	-	0.2	0.0
Quality	0.8	0.4	-	-	-	-	0.2	0.1
Recency	0.3	0.1	0.1	0.0	0.1	0.0	-	-
Scope	0.9	0.4	-	-	-	-	2.2	1.1
User Background	0.2	0.1	0.1	0.0	-	-	1.4	0.6
Format	0.1	0.0	-	-	-	-	-	-
Ranking	-	-	0.1	0.1	-	-	-	-
Resource type	0.1	0.1	-	-	-	-	0.5	0.2
Serendipity	-	-	0.1	0.0	0.1	0.0	-	-
Topicality	3.2	1.1	0.6	0.2	0.1	0.0	1.2	0.5
Tangibility	0.8	0.3	-	-	-	-	0.5	0.1

Table 4. Fixation data across relevance judgments (mean values)

Some interesting findings were also observed in the case of the *individual fixation length* and *FixCriterion* measures. At the *individual fixation length* level the results showed that participants made less but longer fixations on information related to the criterion of *Quality*, mean of 0.53 secs, than

the criteria of *Topicality*, 0.3 secs, (for selected surrogates judged to be very relevant) and *Scope*, 0.51 secs, (for not selected surrogates). Finally, although the criterion of *Affectiveness* occurred less frequently than the criterion of *Scope* when participants decided not to click on a surrogate, the average fixation time spent each time the criterion of Affectiveness occurred was longer (1.2 secs) than the average time spent for each individual occurrence of the criterion of *Scope* (1.1 secs). However, these differences were not statistically significant (one-way within subjects ANOVA).

DICUSSSION

Relevance criteria use and eye movements

The findings of this study confirmed the multidimensional nature of relevance criteria use and especially the fact that participants were not limited to the criterion of topicality when judging relevance. This finding is in accordance with the results of traditional studies on relevance criteria use where researchers found that *Topicality* was not the only criterion used for relevance judgment both in the context of bibliographic information retrieval systems (Barry and Schamber, 1998; Wang and Soergel, 1998) and web search engines (Crystal and Greenberg, 2006; Rieh, 2002). Although most of the criteria used by the participants in our study were similar to the criteria used by participants in other user studies (e.g. Barry and Schamber, 1998), the criteria of *Serendipity* and *Ranking* had not been observed before. Although these were among the least frequently occurred criteria, they both provide an interesting dimension of human predictive relevance judgment behavior in the context of search result interfaces. *Serendipity* was purposively used by participants not able to find *very relevant* or *partially relevant* information after they had clicked on the Top-3 results. The analysis showed that serendipitous judgments were expressed in the case of surrogates ranked at the middle of the first result page (fourth and sixth surrogate). In an earlier study, Foster and Ford (2003) found that serendipity could be regarded as a purposive component of the information seeking process. In addition Crystal and Greenberg (2006) in their study of relevance criteria use found that serendipity was part of some participants' relevance judgment behavior. However, the researchers did not include serendipity in the list of relevance criteria used. *Ranking* was another purposive criterion which demonstrates people's blind trust for Google's ranking algorithm (Pan et al., 2007). However, it is worth mentioning, that the criterion of *Ranking* occurred very rarely among participants and it was expressed by a very short fixation on the summary of the first ranked surrogate (average of 0.12 secs) and a mouse-click on the hyperlinked title of the same surrogate.

Moreover, the findings showed that participants in the study made more and longer fixations on information associated with the criterion of *Topicality*. This finding is in accordance with the results of traditional studies on the use of relevance criteria where researchers found that *Topicality* was the most popular relevance criterion among participants (e.g. Barry and Schamber, 1998; Crystal and Greenberg, 2006). However, the analysis of the *individual fixation length* measure showed that the fixations made on information related to the criterion of *Topicality* were more, but shorter in average length, than the fixations made in the case of the criterion of *Scope*. Similar, were the findings in the case of the *FixCriterion* measure. Since an increase in the length of individual fixations can be associated to an increase in cognitive load (Rayner, 2009), it could be hypothesized that information related to the criterion of *Scope* increased the cognitive effort spent by participants during the evaluation of the search results. However, from the data collected it is not clear whether the increase in cognitive load was related to difficulties in information processing, such as text comprehension, or the importance of the information sought for relevance judgment. In other cases, the *average length of individual fixations* for the criterion of *Topicality* was equal to the length of other criteria, such as *Quality*, *Recency*, *Format* and *User Background* (Table 2). Thus, the findings, of this study suggest that the frequency of occurrence cannot be used as the only measure of the importance of a particular criterion for relevance judgment. Fixation data, such as the number of fixations, total fixation length (*FixLength*), length of individual fixations and *FixCriterion*, along with traditional measures of relevance criteria use (such as the frequency of occurence), can provide a more accurate view of human relevance judgment behavior in the web.

Relevance criteria and ranking order

Also, the observation of participants' eye movements during the process of predictive judgment showed that the use of relevance criteria is dynamic and can be influenced by the ranking order, the different surrogate components (Title, Summary, URL) and participants' relevance judgments (very relevant, partially relevant and not relevant). For example, the number and length of fixations changed as participants moved from the top to the bottom of the result list (Figure 2). In this manner, the amount of attention spent on information related to the criterion of *Topicality* decreased from top to bottom, while other criteria, such as *Scope*, *Tangibility* and *User Background* were fixated more often and for a longer period of time than *Topicality* between the third and the eighteenth ranked result. Although other researchers have observed changes in the use of relevance criteria between information searching stages (Zhang and Yuelin, 2008) and relevance judgment phases (predictive and evaluative judgments) (Tang and Solomon, 2001), this is the first study to show an effect of ranking order on the dynamics of relevance criteria use.

Relevance criteria and surrogate components

Similar changes in participants' fixations were observed in the case of surrogate components. For example, information related to the criteria of *Topicality* and *Scope* was fixated more often and for a longer period of time in the case of the Title component rather than the URL or the Summary.

Similarly, information related to the criteria of User background and Quality tended to be fixated more often and for longer time in the URL component rather than the Title or the Summary (Table 3). Therefore, the amount of attention spent on the different surrogate components can be also associated with the type of criteria used by searchers to judge relevance as well as to other factors, as suggested by previous eye-tracking research, like the design of the search result interface (e.g. list or clustered) (Rele and Duchowski, 2005) or the length of surrogate's summaries/snippets (Cutrell and Guan, 2007).

Relevance criteria and relevance judgments

The examination of participants' relevance judgments (e.g. very relevant, partially relevant, not relevant) showed an effect of relevance criteria use on participants' fixations. For example, fixations on information related to the criterion of *Topicality* were significantly longer than the fixations made for other types of criteria in the case of *very relevant* and *partially relevant* surrogates (Table 4). However, participants made more and longer fixations for information related to the criterion of *Scope* than the criterion of *Topicality* for not relevant surrogates (not selected surrogates). This latter finding aligns with the results of another study, which showed that *Scope* was the most frequently used criterion for the identification of not useful web pages (Tombros et al, 2005). It is worth mentioning that the latter study by Tombros et al was focused on the evaluation of web-pages as opposed to search result interfaces and did not make use of eye tracking technology.

CONCLUSIONS

The purpose of this paper was to present some findings that could shed light on the process of predictive relevance judgment and especially on the problem of relevance criteria use through the use of eye tracking. 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. Tombros et al., 2005; Barry, 1998). However, the present study goes a step further by associating the use of relevance criteria to behavioral 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 studies which made use of eye tracking as the only data collection technique (e.g. Loringo et al., 2006) and were focused on the performance of controlled and simulated search tasks (e.g. Cuttrell and Guan, 2007a; Pan et al., 2007), the present research validates the suitability of using eye tracking in a multi-method research design that involved the performance of real tasks.

The findings of this research could have both theoretical and practical implications in the field of information retrieval and relevance judgment behavior in particular. From a theoretical point of view, the results extend our knowledge about human relevance judgment behavior in the web and suggest the presence of some new fixation patterns as far as concerns the use of certain relevance criteria. For example, it appears that there is an effect of ranking order on the length of fixations on information related to different relevance criteria. In many cases, for surrogates displayed below the Top-3 results, participants were not limited to a trivial judgment of topicality but spent more time fixating on information related to other subjective criteria, such as *Scope*, *Tangibility* and *User background*. Furthermore, the results showed that different surrogate components (Title, Summary and URL) seem to be associated with fixations related to different types of relevance criteria (Table 3). Another finding not observed before refers to the use of relevance criteria in predictive relevance judgment decisions. It appears that users tend to spend more and longer fixation on information related to the criterion of *Topicality* for surrogates judged as very relevant or partially relevant. However, fixation patterns change in the case of not relevant surrogates.

From a practical point of view the results of this type of research could inform the design of information retrieval systems. For example, the emergence of different fixation patterns could be captured for the development of implicit feedback mechanisms and the dynamic display of search result interfaces. The understanding of human visual searching behaviour during relevance judgment could provide a new ground for experimentation in information retrieval, like data log and click-through analysis did in the past. Although this paper presented some novel findings regarding human relevance judgment behavior, further research is needed to test their generalisability using different types of participants (e.g. novice users of the web) and search result interfaces (e.g. bibliographic databases or visual search engines). Finally, besides the findings reported on this paper, 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.

ACKNOWLEDGMENTS

This research is funded by the AHRC.

REFERENCES

Aula, A., Majaranta, P., & Raiha, K-J. (2005). Eye-tracking reveals the personal styles for search result evaluation. In M.F. Constabile & F. Paterno (Eds.), INTERACT 2005 (pp. 1058-1061). Berlin: Springer-Verlag.

Barry, C. (1994). User-defined relevance criteria: an exploratory study, *Journal of the American Society for Information Science and Technology*, 45 (3) 149-159.

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

Barry, C. and Schamber, L. (1998). Users' criteria for relevance evaluation: a cross-situational comparison. *Information Processing and Management*, 34 219-236.

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.

Cutrell, E., & Guan, Z. (2007). What are you looking for? An eye-tracking study of information usage in web search. In Proceddings of the ACM HCI 2007 Conference on Human factors in computing systems (pp. 407-415). New York: ACM.

Dumais, S., Buscher, G. & Cutrell, E. (2010). Individual differences in gaze patterns for web search. *IIiX 2010*, August 18-21, New Brunswick, New Jersey USA.

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

Foster, A. & Ford, N. (2003). Serendipity and information seeking: an empirical study. Journal of Documentation, 59, 321-340.

Goldberg, J. H., Stimson, M., Lewenstein, M., Scott, N., & Wichansky, A. (2002). Eye tracking in web search tasks: design implications. In Proceedings of the Eye tracking Research and Applications Symposium (ETRA). (pp. 51-58). New York: ACM.

Granka, L., Joackinms, T. and Gay, G. (2004). Eye-Tracking Analysis of User Behavior in WWW Search. In: M. Sanderson, K. Jarvelin, J. Allan, P. Bruza (Eds.), *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield , July 25-29, 2004*. (pp. 478-479). New York: ACM.

Hughes, A., Wilkens, T., Wildemuth, B., and Marchionini, G. (2003). Text or pictures? An eyetracking study of how people view digital video surrogates. In E.M. Bakker et al (Eds.), Proceedings of the Second International Conference, CIVR 2003 Urbana-Champaign, IL, USA, July 24–25, 2003 (pp.271-280) Berlin: Springer-Verlag.

Loringo, L. (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.

Loringo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence of task and gender on search and evaluation behavior using Google. Informaiton processing and management 42, 1123-1131.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G. and Granka, L. (2007). In Google we trust: users' decisions on rank, position and relevance. *Journal of computer-mediated communication*, 12 801-823.

Papaeconomou, C., Zijlema, A., & Ingwersen, P. (2008). Searchers' Relevance Jugements and Criteria in Evaluating Web Pages in a Learning Style Perspective. In *IIiX 2008*, London, UK, 123-132.

Pirolli, P. (2007). *Information foraging theory: adaptive interaction with information*. Oxford: Oxford University.

Rayner, K. (2009). Eye movements and attention in reading, scene perception and visual search. *The quarterly journal of experimental psychology*, 62, 1457-1506.

Rele.S., & Duchowski, A.T. (2005). Using eye tracking to evaluate alternative search results interfaces"Impact" of journals as proxies for citation counts. In *Proceedings of the 49th annual meeting of Human Factors and Ergonomics society, September 26-30 2005, Orlando, FL, HFES*. Orlando, Florida, HFES, 2005.

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.

Rodden, K., Fu, X., Aula, A. & Spiro, I. (2008). Eye-mouse coordination patterns on web search pages. In Proceedings of the ACM CHI 2005 conference on human factors in computing systems (pp. 2997-3002). New York: ACM press.

Spink, A. (2002). A user-centered approach to evaluating human interaction with web search engines: an exploratory study. *Information processing and management*, vol. 38, pp.401-426.

Tang, R. and Solomon, P. (2001) Use of relevance criteria across stages of document evaluation: on the complementarity of experimental and naturalistic studies, *Journal of the American Society for Information Science and Technology*, 52 (8) 676-685.

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.

Tseng, Y-C., & Howes, A. (2008). The adaptation of visual search strategy to expected information gain. In Proceedings of the ACM CHI 2008 Conference on human factors in computing systems (pp. 1075-1084). New York: ACM press.

Wang, P. and Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science and Technology*, 49(2) 115 – 133.

Zhang, Y. and Yuelin, L. (2008). A user-centered functional metadata evaluation of moving image collections. *Journal of the American Society for Information Science and Technology*, 59(8) 1331-1346.