A User Study of Web Search Session Behaviour using Eye Tracking data

Mari-Carmen Marcos
Dept. Communication,
Universitat Pompeu Fabra,
Barcelona, Spain.
mcarmen.marcos@upf.edu

David F. Nettleton[1,2]
[1]Dept. Information Technology
and Communications,
Universitat Pompeu Fabra,
Barcelona, Spain.
[2]IIIA-CSIC, Bellaterra, Spain.
david.nettleton@upf.edu

Diego Sáez-Trumper
Dept. Information Technology
and Communications,
Universitat Pompeu Fabra,
Barcelona, Spain.
diego.saez@upf.edu

Abstract. In this paper we present and empirically evaluate a user study using a web search log and eye tracking to measure user behaviour during a query session, that is, a sequence of user queries, results page views and content page views, in order to find a specific piece of information. We evaluate different tasks, in terms of those who found the correct information, and in terms of the query session sequence itself, ordered by SERP (Search Engine Result Page), number and return visits to the results page for the same query. From this we are able to identify a number of different behaviour patterns for successful and unsuccessful users, and different trends in user activity during the query session. We find that a user behaves differently after the first query formulation, when we compare with the second formulation (both queries being for the same information item). The results can be used to improve the user experience in the query session, by recognising when the user is displaying one of the patterns we have found to have a low success rate, and offering contextual help at that point. The results may also contribute to improving the design of the results page.

Keywords: Web search, eye tracking, query session, user behaviour, search user experience.

1. INTRODUCTION

Public domain search engines are one of the most utilized tools on the Internet for accessing online information. However, when users are unable to find what they are looking for, they feel frustrated and this affects their user experience. These reasons, among others, are the motivation for incorporating changes in the search engine's functionality and interface which will enhance user support.

The main source that search engines have to know what the users are doing in the results and contents pages are weblogs, from which a wide variety of information can be obtained. For example, typical descriptive variables include: the terms that were used to formulate a query, how much time was spent in the SERP (Search Engine Results Page), which result(s) were selected, how long the user took to return to the SERP from the content page (if they return), whether the original query was reformulated or another result selected, and how long they took to do this. However, although much work has been done on search engine usage, it has mainly focused on the analysis of large anonymous web search logs, or on specific user studies which analyse the user's reaction to specific items on the screen (snippets, publicity banners, predefined regions, etc.). In this work we offer a novel approach, in which we study the user behaviour in the context of a query session whose substructure consists of different SERP views and query formulations in order to find a given piece of information. The data captured and made available for analysis consists of the web search log data (number of queries formulated, number of results clicked, time duration on page, …), together with the data which is obtained by an Eye Tracking device, which enables us to obtain statistics about ocular activity, such as fixation rate and fixation duration.

Thus, in this paper we derive two aspects of utility for Human-Computer Interaction in SERP design and online help: **(i)** a behaviour model which classifies the user with respect to their navigation pattern; **(ii)** key trends of specific search log and eye tracker variables which have been shown to be statistically significant.

© The Authors. Published by BISL.

Proceedings of the BCS HCI 2012
People & Computers XXVI, Birmingham, UK

The structure of the paper is as follows: in Section 2 we briefly detail the related work; in Section 3 we describe the experimental design and methodology; in Section 4 we present the empirical results: (i) a user search behaviour model, and (ii) an analysis of the trends shown by the different patterns in terms of the descriptive variables and task success; finally, in Section 5 we make some conclusions and propose future work.

2. RELATED WORK

In recent years, web search behaviour has been analyzed in detail by different research groups in the web mining community using query logs collected by search engines (Baeza-Yates, Hurtado, Mendoza and Dupret, 2005), (Nettleton, Baeza-Yates, 2008) and (Nettleton and Codina-Filba, 2010). In (Nettleton and Codina, 2010) a query session model was proposed which represented a sequence of query formulations and reformulations incorporating a cost function defined in terms of the available log variables. On the other hand, the technology of Eye tracking has been applied mainly in ergonomic user studies and studies of how the user behaves with respect to specific objects or characteristics within an image on the computer screen or a web page (Dogusoy and Cagiltay, 2007) and (Pan, Hembrooke, Gay, Granka, Feusner and Newman, 2004).

References in the literature which relate behaviour in a complete query session to the SERP (our current work) are scarce. However, there exists a greater diversity of publications of more general studies relating to different aspects of behaviour in a given SERP. Some studies have been carried out related to the distribution of the gaze on different areas of the SERPs (González-Caro and Marcos, 2011)(Guan and Cutrell, 2007) and (Lorigo, Pan, Hembrooke, Joachims, Granka and Gay, 2006). Other studies, such as (Baeza-Yates and Castillo, 2001) and (Buscher, Dumais and Cutrell, 2010), have differentiated between organic results and publicity, and made distinctions in terms of the type of intention which motivates a query depending on if the search task is informational, navigational (Lorigo, Pan, Hembrooke, Joachims, Granka and Gay, 2006) or transactional (González-Caro and Marcos, 2011).

There is less work applying Eye Tracking to query session analysis in search engines, which provides the motivation for the current work. Some examples of work in this area are (Cutrell and Guan, 2007), (Goldberg, Stimson, Lewenstein, Scott and Wichansky, 2002) and (Pan, Hembrooke, Joachims, Lorigo, Gay and Granka, 2007), which tend to focus on specific „areas of interest' of the screen content, such as „snippet length'. In contrast to these studies, in the present work we allow users to formulate their own queries based on a given search objective. Our work also covers a gap in the literature with respect to the evaluation of user behaviour within a query session, whose substructure consists of different SERP views and query formulations.

3. EXPERIMENTAL DESIGN AND METHODOLOGY

In order to model the user we use data from the web search log together with data which is specific to ocular activity (gaze). The experiments have been conducted in a lab and are designed to simulate a natural web search session in which users can formulate free text queries, look at the returned documents, click on links and visualize the corresponding documents, formulate a new query, click on the new results, and so on. The users are initially given a specific task and they have to search for the information to resolve that task.

Three general tasks were defined in a non specialized topic and with a unique correct answer. In this way we to reduce the ambiguity of the results and give the users an equal a priori chance of finding the information. Each task is independent and different from the other question.

(i) T1: "Name of a mechanical machine (not electrical) for calculating, of German origin, which fitted in the palm of a hand": The correct answer to the question is "Curta"
(ii) T2: "Name of the wife of the author of "The Jungle Book" ": The correct answer to the question is "Carrie Balestier"
(iii) T3: "Name of a Catalan NGO which works in India and whose founder was recently hospitalized": The correct answer to the question is: "Vicente Ferrer Foundation"

We note that we took measures in order to avoid bias and balance the user sessions. To do this we randomized the assignment and the order of the tasks to the users. Another key aspect was the design of the tasks: we designed what we considered to be a "difficult" task (i), a task of "average" difficulty (ii), and a relatively "easy" task (iii). Task (iii) was relatively easy because at the time of the experiments there was a story in the media about the hospitalization of Vicente Ferrer. In order to establish whether a task was successfully completed, at the end of each task, the moderator labelled it as being successful or unsuccessful.

There were total of 57 participants with ages ranging from 18 to 61 years and whose average age was 27. Students, lecturers and administration personnel were recruited from the Universitat Pompeu Fabra and Barcelona Media, selected so

as to give equilibrium between age, gender and education levels. All participants had an average level of competence for searching in Internet, (defined as users who searched the Web for information at least once a week). None of the users had previously used an eye-tracker. The hardware used for measuring eye movements was the Tobii 1750 Eye Tracker (Ewing, 2005).

Each user was asked to perform two search tasks, corresponding to two of the three defined questions. The user had to find the correct answer to the question by searching the web using Google. The users could freely click links and scroll page-up/page-down as needed, in order to approximate as much as possible to a real Internet search, with a 3 minute time limit for practical reasons.

The results were analysed in terms of three different types of variables. The first type were based on session log data obtained from the eye tracker: timestamp recording of the session duration measured in milliseconds; the number of SERPs seen by a user in a session for solving a task; number of content pages a user visits in a session for solving a task; number of terms typed in the first two queries of the session. The second type of variables were based on gaze behaviour data obtained from the eye tracker, namely, the number of fixations by SERP and the fixation duration by SERP. Finally, the third type of variable was the task performance noted by the moderator at the end of each task, labelled as successful or unsuccessful.

In order to be clear about what we mean by a query session, consider a user who is given task (ii), "name of the wife of the author of "The Jungle Book". The user formulates a first query "the jungle book author" in order to find the name of the author. The user looks at the results page but does not select any documents. In a second query the user formulates "jungle book Kipling wife", that is, having found the surname of the author, it is used to make the new query more specific. The user selects and reads a results document. Finally, in a third query the user formulates "Rudyard Kipling wife" which presents the best results page and the user finds the answer in one of the results documents. Together, this represents a session consisting of three different queries, three different SERP views and two document selections.

As this is an experimental study, we have to consider up to what point it is representative of reality. In terms of real-life search engine usage, we have tried to achieve this by defining informational queries of a type which users often tend to carry out. The questions were non trivial, given that all three almost always required refinement in order to obtain the answer (e.g. „Name of the wife of the author of the Jungle Book' are two questions in one).

4. RESULTS

In this section we first present the user search behaviour model and explain how it was elicited from the data log. Then we analyze different facets of the data: search success with respect to pattern, the relation of the available variables with search success and the visual behaviour on the SERPs.

4.1 User search behaviour model

First we synthesized the raw historical data log generated by the Eye Tracker in order to obtain a summarized log an excerpt of which can be seen in Table 1. In Table 1 we see in the first column the task id, in the second column the user id and in the third column if the query session was successful. Then, columns 4,5 and 6 are used to identify the user sequence: sequence, query and visit. With this we can define the pattern (column 8) and hence the pattern code (A to E). For example, the first two rows of data represent a query session for task T1 and user P01, that is, a task/user tuple. In the first row the corresponding query and visit numbers are {1,1} and in the second row {2,1}. Hence, as there are no more lines for this task/user tuple, the resulting pattern will be 11-21, which we designated as pattern C. It can be seen that all the other pattern assignments to query sessions are formulated similarly. Finally, the columns 9 to 14 consist of data of web log and Eye Tracker (ocular) variables. By processing the historical log generated by the Eye Tracker as we have just explained, we were able to identify five principal patterns of user behaviour in query sessions.

In Figure 1 we see a graphic representation of the patterns, designated as A, B, C, D and E. In the following we briefly describe the user behaviour corresponding to each pattern.

Pattern A: users type a query, look at the first SERP and afterwards they navigate to content pages without coming back to the search engine page in any moment of the session. In this pattern, users see just once the search engine page. *Pattern B:* users formulate a query, look at the first SERP, visit one or more content pages, and then return to the search engine page to formulate a new query. Then they visit one or more content pages and terminate the session. So, they see two different SERPs: the first one and a second one with a different list of results because they have refined the query. *Pattern C:* as in pattern B, users formulate a query, look at the first SERP, visit one or more content pages, then return to the search engine page to formulate a new query, but they do not terminate the session: they go again to the list of results on the search engine to choose another result, to refine the query or to formulate a new query. So, they see the search engine page more than twice.

Table 1: *Processed eye tracker data log*

TASK	USER	SUCCESS	SEQUENCIAL	QUERY NUMBER	VISIT NUMBER	Pattern code*	Pattern	Number of serps	Num terms	FIXATIONS NUMBER (#)	FIX ATIONS DURATION (ms)	FIXATIONS PATH (pixels)	FIXATIONS DISPERSION (pixels / fixations number)
T1	P01	**No**	1	1	1	C	11-21-…	6	4	24	11285	2401	100
T1	P01	**No**	2	2	1	C			4	7	2106	556	79
T1	P23	**No**	1	1	1	E	11-12-…	3	3	3	5064	244	81
T1	P23	**No**	2	1	2	E			3	6	4939	197	32
T1	P03	**No**	1	1	1	C	11-21-…	5	2	32	25023	3864	120
T1	P03	**No**	2	2	1	C			7	39	11575	3857	98
T1	P09	**No**	1	1	1	A	11	1	3	20	21076	3511	175
T1	P15	**Yes**	1	1	1	A	11	1	6	8	7945	1398	174
T2	P47	**Yes**	1	1	1	B	11-21.	2	6	13	13647	2762	212
T2	P47	**Yes**	2	2	1	B			2	2	1524	254	127
T1	P20	**No**	1	1	1	D	11-12.	2	5	9	3395	1171	130
T1	P20	**No**	2	1	2	D			5	16	8419	3235	202

*Pattern code: A=11, B=11-21., C=11-21…, D=11-12., E=11-12…

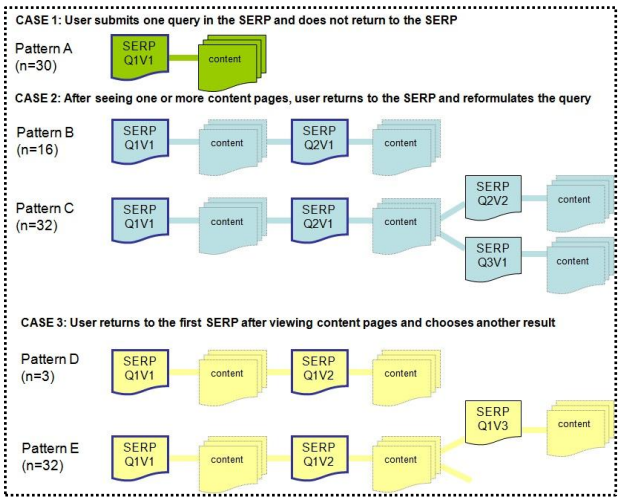

Figure 1: *User navigation search model*

Pattern D: in this case, users write a query, look at the first list of results, visit one or more content pages, and then return to the search engine page to recheck the documents provided by the search engine. Then they visit one or more content pages and terminate the session. So, they see just one list of results, but they see it twice. *Pattern E:* in a similar manner to pattern D, users write a query, look at the first list of results, visit one or more content pages, and then return to the search engine page to recheck the documents provided by the search engine. Then they visit one or more content pages, and return at least one more time to the SERP to choose another result. Thus, in pattern E they see the same list of results of the search engine more than twice before ending the session.

In Figure 1 a schema is shown of the 5 user behaviour models we have identified from the user sessions. In the nomenclature, Q indicates to which query the results correspond (Q1 is the first query, Q2 is a query reformulation, and so on); V indicates the number of times a user saw a given SERP (results list), where V1 signifies the first time, V2 the second time, and so on. Finally, *n* indicates the number of *tuples* user/task for a given pattern. For example, with reference to Figure 1, pattern B has two possible states: SERP Q1V1 and SERP Q2V1, and has *n* = 16. That is, 16 users formulated a first query, saw the first results page just once, then formulated a second query and saw the second results page just once.

4.2 Analysis of the trends shown by the different patterns in terms of the descriptive variables and task success

With reference to Table 2, we found that there were two patterns with a high relative success rate for solving the tasks (patterns A and B with 96% and 81%, respectively) and two other patterns with a lower success rate (patterns C and E with 44% and 29%, respectively). We could not calculate the success rate for pattern D because the sample size was too small. We analyzed a total of 101 cases of user/task *tuples*. In Figure 2 we see the relation between behaviour pattern, the SERP number and the average fixation duration (in milliseconds). We see some distinguishing trends, for example, a much higher fixation duration for SERP 2 and pattern E, and a lower fixation duration for SERP 2 with respect to SERP 1, for patterns B and C.

Table 2: *Success rate by pattern*

	Success / Failure			
Pattern	**A**	**B**	**C**	**E**
%SUCCESS	96.77	81.25	43.75	28.13
%FAIL	3.23	18.75	56.25	71.88
Sample size	31	16	32	32

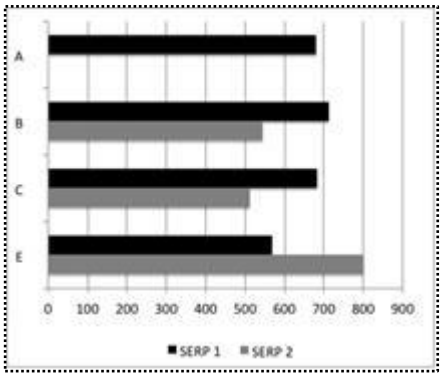

Figure 2: *Average fixation duration (milliseconds) on the first and second SERP for each of the behaviour patterns (A, B, C and E)*

Table 3: *Descriptive Statistics of key variables Vs behaviour patterns (Non ocular variables)*

	Task Duration			
Pattern	**A**	**B**	**C**	**E**
Average	74951	104359	159182	162009
St. Dev.	37522	51994	32766	38347
Sample size	31	16	32	32
	Total Content Pages Visited			
Pattern	**A**	**B**	**C**	**E**
Average	1.55	1.69	2.28	2.69
St. Dev.	0.99	1.01	1.20	1.91
Sample size	31	16	32	32
	Number of SERPS Seen			
Pattern	**A**	**B**	**C**	**E**
Average	1.00	2.00	5.41	5.09
St. Dev.	0.00	0.00	1.90	1.87
Sample size	31	16	32	32

With reference to Tables 3 and 4, we see the results for the descriptive variables, and by pattern, for the non 'ocular' variables (Table 3) and the 'ocular variables' (Table 4). In Table 3 we see the statistics for the variables which gave the strongest correlation with the pattern and their corresponding success rates (Table 2). For example, the most successful patterns (A and B), had significantly lower average durations of 74951 and 104359 ms, respectively, in comparison with the least successful patterns (C and E), whose average durations were 159182 and 162009 ms, respectively. A similar relation is evident for 'number of SERPS' and 'Visited pages'. In Table 4 we again see the variables which showed the strongest relation with the pattern and the success rate (there were two other ocular variables whose correlation was not significant, see last two columns of Table 1).

Table 4: *Descriptive Statistics of key variables Vs behaviour patterns (Ocular variables)*

	Num. of Fixations SERP 1			
Pattern	**A**	**B**	**C**	**E**
Average	15.39	17.94	19.63	19.26
St. Dev.	15.49	17.67	17.57	20.81
Sample size	28	16	27	27
	Num. of Fixations SERP 2			
Pattern	**A**	**B**	**C**	**E**
Average		13.07	18.20	11.89
St. Dev.		9.04	19.44	10.20
Sample size		15	25	27
	Fix. Duration SERP 1			
Pattern	**A**	**B**	**C**	**E**
Average	10480	12771	13402	10914
St. Dev.	11391	12960	12088	7262
Sample size	28	16	27	27
	Fix. Duration SERP 2			
Pattern	**A**	**B**	**C**	**E**
Average		7112	9311	9513
St. Dev.		4488	8373	9033
Sample size		15	25	27

We observe that the variable 'number of fixations SERP 1' shows a general trend in that patterns A and B (the most successful) have lower values than pattern C and E (the least successful). In terms of 'Fix. Duration', if we consider the difference in the average values between 'Fix. Duration SERP 1' and 'Fix. Duration SERP 2', we see that the more successful pattern B manifests a significant fall (from 12771 down to 7112), whereas for the least successful patterns, C manifests a significantly lesser fall (from 13402 down to 9311) and pattern E (from 10914 to 9513).

With respect to the relation between task/task difficulty and behaviour pattern, we would expect that a more difficult task will probably require more query reformulations and page views. With reference to the difficulty levels of the tasks explained in Section 3, T3 (easy) was concentrated in patterns A and B (48% and 50% of all tasks for those patterns, respectively), T2 (average) was fairly evenly distributed throughout patterns, and T1 (difficult) was concentrated in pattern E (66%). Refer to Figure 1 for the corresponding pattern structures. Finally, with respect to statistical significance tests, we performed the two tail *t-test* to obtain the *p-values* for the last viewed page *versus* the previously viewed page, for each pattern and ocular variable (for example, with reference to Table 1, a comparison was made of row 1's data with that of row 2, for user P01 and task T1). However, given the small sample sizes and high standard deviations relative to the average values (see Table 4), the *p-values* were not considered reliable. For example, the two tail *t-test* with respect to pattern and variable gave *p-values* of 0.12 (pattern E with respect to number of fixations), 0.09 (pattern B with respect to fixation duration) and

0.12 (pattern C with respect to fixation duration). As future work, we could perform the *t-test* on other combinations of variables and states.

5. CONCLUSIONS

Following the observation and the analysis of the logs we have proposed a model of user search behaviour which consists of 5 possible navigation patterns, in terms of the route that the users follow from the first results page through to the end of the session. The descriptive variables consist of non ocular data (web log) together with ocular data (eye tracker).

We have found that the users who correspond to behaviour patterns C and E have the lowest success rates and therefore are those who need the most help/support in order to be successful in their searches. We note that patterns C and E account for approx. 50% of the total cases. What differentiates these two patterns from the others is that the users visit the same search results page multiple times. On the other hand, those of pattern A and B only visit the first SERP once (indicated by V1). Current search engines do not generally have the functionality to guide the user in their search strategy, instead the SERP page remains static waiting for the user to choose another contents page or to reformulate their query. For this reason, the second SERP view seems to be the best moment for the search engine to take the initiative and try to accompany the user, reorient him/her, and show elements in the interface which help him/her to find the right search path.

As future work we hope to develop a help interface (api) for the SERP page, based on the detected behaviour patterns. Then, using this interface we will be able to conduct tests and statistical evaluation on a larger population of users, in order to evaluate how the model generalizes and use the results to give feedback for any necessary model redesign and tuning.

6. ACKNOWLEDGEMENTS

This research is partially supported by the Spanish MEC, (project HIPERGRAPH TIN2009-14560-C03-01).

7. REFERENCES

Baeza-Yates, R., Hurtado, C., Mendoza, M. and Dupret G., (2005). Modeling user search behavior. In Proc.3rd Latin American Web Congress, 31 Oct. -2 Nov, Buenos Aires, Argentina, p. 242 – 251, IEEE Washington DC, USA.

Buscher, G., Dumais, S., and Cutrell, E. (2010). The good, the bad and the random: An eye-tracking study of ad quality in Web search. In Proc. SIGIR 2010, 19-23 July, Geneva, Switzerland, pp. 42–49, ACM, NY, USA.

Cutrell, E. and Guan, Z., (2007). What are you looking for? An eye-tracking study of information usage in Web search. In Proc. CHI 2007, 30 April - 03 May, San Jose, CA, USA, pp. 407 – 416, ACM, NY, USA.

Dogusoy, B. and Cagiltay, K. (2007). Evaluation of a Visually Categorized Search Engine. In Proc. 3rd Technology-Enhanced Learning Enlargement Workshop, September 2007, Sofia, Bulgaria, pp. 20-30, Bulgarian Sociological Association.

Ewing, K. (2005) Studying Web Pages Using Eye Tracking, Tobii Technology, http://www.scribd.com/doc/20321480/Tobii-Whitepaper-Studying-Web-Pages-Using-Eye-Tracking (1st August 2012).

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., and Wichansky, A. M., (2002). Eye tracking in web search tasks: design implications. In Proc. ETRA 2002, 25-27 March, New Orleans, USA, pp. 51-58, ACM, NY, USA.

González-Caro, C. and Marcos, MC. (2011). Different Users and Intents: An Eye-tracking Analysis of Web Search. In Proc. WSDM 2011, 9-12 February, Kowloon, Hong Kong, pp. 1-8, ACM, NY, USA.

Guan, Z. and Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In Proc. CHI 2007, 28 April-3 May, San Jose, California, USA, pp. 417–420, ACM, NY, USA.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., and Gay, G. (2006). The influence of task and gender on search evaluation and behavior using Google. IP&M, 42(4): 1123–1131.

Nettleton, D.F. and Baeza-Yates, R. (2008), Web retrieval: Techniques for the aggregation and selection of queries and answers. *Int. Journal of Intelligent Systems*, 23(12), 1223-1234.

Nettleton, D.F. and Codina-Filba, J. (2010), "A Cost-Continuity Model for Web Search". *Springer LNAI*, 6408, 219-230.

Pan, B., Hembrooke, H. A., Gay, G.K., Granka, L.A., Feusner, M.K. and Newman, J. K., (2004). The Determinants of Web Page Viewing Behavior: An Eye-Tracking Study. In Proc. 2004 Symp. on Eye tracking research & applications. 22-24 March, San Antonio, Texas, USA, pp. 147 – 154, ACM NY, USA.

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(3), article 3. http://jcmc.indiana.edu/vol12/issue3/pan.html (1st August 2012).