Browsing Patterns in Retrieved Documents

Jaana Kekäläinen
Tampere University
33014 Tampere
Finland
jaana.kekalainen@uta.fi

Paavo Arvola
Tampere University
33014 Tampere
Finland
paavo.arvola@uta.fi

Sanna Kumpulainen
Tampere University
33014 Tampere
Finland
sanna.kumpulainen@uta.fi

ABSTRACT
The paper reports a test exploring how retrieved documents are browsed. The access point to the documents was varied – starting either from the beginning of the document or from the point where relevant information is located – to find out how much browsing and context the users need to judge relevance. Test results reveal different within-document browsing patterns.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query formulation, Search process.*

General Terms
Experimentation, Human Factors.

Keywords
Document browsing patterns, best entry point.

1. INTRODUCTION
Locating relevant parts within a document is important in factual searches and question answering where users are presumably looking for focused information. However, the interpretation of information may depend on its context. The user may need more than the exact fact or answer to make sense of the information, and to judge the reliability or relevance of the source. [4]

A typical information retrieval (IR) system offers a result list to the best matching documents for the users. Standard document access delivers the beginning of a document to the screen and leaves the task of locating relevant content for the searcher. In this study, we test a more focused approach: instead of document beginning, we automatically set the screen state to the point of the document where the relevant information is immediately accessible (*best entry point*). Our research question is: *Are the users able to identify the relevant information at the best entry point or do they need more context for their relevance judgment?* We executed a laboratory experiment comparing two different access modes to the retrieved documents, the beginning of the document and the best entry point. We report the results of a pilot study.

IIiX '14, August 26 - 29 2014, Regensburg, Germany

ACM 978-1-4503-2976-7/14/08…$15.00.
http://dx.doi.org/10.1145/2637002.2637047.

2. RELATED WORK
User's browsing within documents and their need for textual context while judging the relevance of documents are focal issues for the present study. Lozides and Buchanan [5] studied users' navigation within documents during document triage for two information seeking tasks. The aim was to explore which parts of the retrieved documents the users view. Users' navigation in PDF files was logged with viewing time for the contents visible on the screen. The researchers suggest four navigational patterns: 1) *step navigation*, where the users move stepwise from the beginning of the document to the end; 2) *flatline*, where the users view only the first visible part of the document; 3) *mountain*, where the users move stepwise from the beginning of the document to the end, and then back to the beginning again; 4) *begin and end*, where the users view the beginning of the document and then quickly scroll to the end of the document without viewing the middle part. The researchers point out that the patterns are nominal perfect models, and the users' actual patterns are combinations of these.

Lin and others [4] studied what size of textual context for answers the users preferred in a question answering system. Of the four alternatives given (exact answer, answer-in-sentence, answer-in-paragraph, answer-in-document) the paragraph size was the most popular and the exact answer without context the least popular.

The idea of the best entry point was introduced by Lalmas and Reid [1]. Focused retrieval with the best entry point is defined as follows [7]: "*The information system presents the user with an article, pre-scrolled to the right location so that the snippet is on-screen in context, for the user to read.*" Here the snippet refers to relevant information rather than a summary of a document. It seems that focused retrieval serves best factual or precise searches, and searching long documents [6].

There are different ways to represent a focused retrieval result: the relevant part may be shown in the surrounding textual context or in the context of the whole document [5]; the relevant part of the document may be extracted from the document and returned as standalone retrieval result [3]. A simple solution is to direct the searcher to the best entry point in the document, as proposed above. This best entry point should be related to the searcher's query and the searcher should be able to identify the information searched for.

3. DATA AND METHODS
A laboratory test was executed to compare settings where the documents are shown to the user (a) starting from the beginning of the document or (b) starting from the best entry point. The idea is to give the participant a factual search task, then log his search session, and finally log the browsing of the first opened document which would contain the relevant information. The retrieved document was represented to the participant either in the mode (a), called document mode (D for short), or in the mode (b) called best entry point mode (BEP). In the current test setting it was not possible to log browsing of any deliberate retrieved document, we

thus guided the searcher to a specific web page (target document). All tasks were performed in the same browser (Firefox®) with the same search engine (Bing®); the access mode to the target document was controlled.

3.1 Participants, tasks and target documents

The participants were recruited from an optional university course at the bachelor's level. The subject of the course is experimental information retrieval. Altogether seven males and seven females participated in the test. The major study subjects of the participants were computer science (7 students), information studies (6 students) and communication studies (1 student). No remuneration was given to the participants.

Factual search tasks were assigned to the participants (see Appendix). Task 1 was for orientation and Tasks 2-4 for actual testing. The topics of the tasks were related to laws and regulations, and to a known researcher's work. Tasks 1, 2 and 4 were rather general, but Task 3 was tailored for these students. The task descriptions were all in the participants' native language. The only exception was the name of the researcher's model, which was given in English (Information Seeking Process by C. Kuhlthau in Task 3).

The target documents encompassed the information required in the search tasks. The selection criteria for the target documents were length and subject: pages should be long enough, in other words, all relevant content should not be visible without scrolling; the subject of the page should suit task construction. The target documents for the test tasks were three web pages: a web page representing the law concerning apartment renting (legal text, Task 2), a web page representing the tax declaration procedure (legal text, Task 4), and Carol Kuhlthau's homepage about her information seeking process model (Task 3). The target documents for Tasks 2 and 4 were in the participants' native language; the target document for Task 3 was in English.

The lengths of the target documents were 8,190 words for Task 2, 5,372 words for Task 3, and 14,445 for Task 4. The entry points of the BEP mode were the places where the relevant information began. The relevant information was only to be found at one place in each document and the point was selected by the researchers. The best entry points ley at different places with regard to the document length: the entry point of the target document for Task 2 was after 443 words (~5% of the text); the entry point of the target document for Task 3 was after 4394 words (~82% of the text); the entry point of the target document for Task 4 was after 1095 words (~7% of the text).

3.2 Test setting

Two test groups were formed: both received search results in the BEP and D mode but the order of the modes was reversed in the groups. The test was run for one group at time in a computer class room. The test protocol included a pre-tasks questionnaire, four search tasks and a post-task questionnaire after each task. The participants had ten minutes to finish each task and the questionnaire. Each test session was video recorded, but only half of the participants were observed because of technical issues.

The pre-tasks questionnaire was about the participants' demographic characteristics and basic searching behavior. The post-task questionnaire elicited the participants' opinions about familiarity with the topic, easiness and realism of the task, and participants' self-evaluation of their success in accomplishing the task. The questionnaire had five statements with four response alternatives (strongly agree, somewhat agree, somewhat disagree, strongly disagree). Sixth question was to control whether the participants noticed anything unusual (yes, no) about the results and that was followed with an open question about what was unusual. This was because in BEP mode the target document did not start from the beginning, and because the target document did not necessarily correspond to the clicked search result entry.

The participants were asked to abandon querying after they had chosen and clicked any retrieved document. That was because all participants were directed to the same, task specific target document irrespective of the result list – the web page we were observing. This makes the querying somewhat unrealistic but we wanted to catch the first query before interaction with the retrieved documents. The participants were asked to write down the relevant information required in the written task description.

The first task (Task 1 in Appendix) was for orientation and its search results were not manipulated. After that, Tasks 2-4 were given to the participants one by one. In these tasks the users were guided to the target document irrespective of what they clicked in the result list. For technical reasons the order of the tasks was same in both groups. In Group 1 the result page for Tasks 2-3 was represented in D mode and for Task 4 in BEP mode. In Group 2 the modes were reversed, i.e. BEP for Tasks 2-3 and D for Task 4.

Queries were logged using a proxy server (Apache®). For logging the within document browsing we used UtaProxy software [2]. UtaProxy is a tool for observing user's browsing within a web document, including mouse actions, scrolling and searching.

3.3 Data analysis

The groups are not balanced because two initially recruited participants did not turn up. Further, the number of participants differs in different tasks because two participants did not finish all the tasks in Group 2. In Group 1 there were 8 participants, and in Group 2 the number of participants was 6. The number of tasks performed in D mode is 20, and the number of task performed in BEP mode is 19. For these reasons we do not emphasize the quantitative analysis but rather try to illuminate the phenomenon in a qualitative manner. No statistical tests were applied because of the small number of participants. The results are indicative, yet showing interesting trends.

4. PRELIMINARY RESULTS

In general the tasks were familiar to the participants but there is variation between tasks (see Table 1). Tasks 1 (orientation) and 3 were found less easy and less interesting than the two other tasks. The participants were also less confident of their success in these tasks, especially in Task 3. The participants did not find the results unusual and only few of their comments (four out of 23) mentioned that the clicked page was not the one they expected (sixth statement in Table 1).

The correctness of the answers was not emphasized in the test but the participants were very determined to get relevant documents to the result list. As they were briefed not to return to querying after clicking at any retrieved document, they did relevant judgments on the basis of the snippets and modified their queries accordingly. They wrote several queries per task (averages per tasks 6, 5, 4, 6; range from 1 to 23). In general, the participants found it disturbing not to be allowed to interact with the whole documents during querying.

The number of correct answers is indicative of the perceived difficulty of the task (see Tables 1 and 2). There seems to be no difference in BEP versus D mode regarding the number of correct answers. In Task 3 there are few correct answers is in both modes. The reason for the perceived difficulty and poor performance is

Table 1. Post-task questionnaire results, averages and medians (in brackets)

Statement	Task # 1 N=13†	2 N=13	3 N=13	4 N=12
The topic was familiar*	3.5 (4)	2.2 (2)	2.4 (2)	2.8 (2)
I have searched similar topics before*	3.5 (4)	2.3 (2)	2.8 (3)	2.6 (2.5)
The task was interesting*	2.5 (2)	1.8 (2)	2.2 (2)	1.8 (2)
The task was easy*	2.4 (2)	2.2 (2)	2.8 (3)	2.4 (2)
I finished the task successfully*	1.8 (1)	1.5 (1)	2.4 (2)	1.7 (1)
The results differed from usual**	1.2 (1)	1.5 (2)	1.7 (2)	1.5 (1)

* scale 1=strongly agree; 2=somewhat agree; 3=somewhat disagree; 4=strongly disagree

** scale 1=no; 2=yes

† N is not 14 in because in every task someone either did not answer all questions or did not complete the task.

obviously that the participants did not understand the goal of the task (poor wording) and perhaps also the foreign language of the target document.

4.1 Browsing sessions

There is a slight indication that browsing times in BEP mode were shorter than in D mode but these results should be taken with a grain of salt since the groups were small and not balanced (BEP averages – min:sec – in Tasks 3-4: 2:55/3:26/2:22; D averages: 5:31/6:46/2:45).

More interesting than the times spent are the browsing patterns the participants utilize. Figure 1 depicts the browsing sessions of three participants in Tasks 2-3. These sessions are selected to illustrate browsing. Although they are examples, some trends emerge.

The rows in Figure 1 represent Tasks 2 and 3 respectively. Browsing time is the length of the browsing session. The y-axis visualizes the whole document length (in percentages) and x-axis denotes the passing of time. The light pink[i] area shows the part of the document that is visible to the participant and also how browsing of the document proceeds over time. The strong black lines show the visibility of the best entry points. Because the target document for Task 2 is longer than the target document for Task 3, the visible area in respective figures is thinner. Very rapid scrolling shows as a very thin pink line. Another way to move fast around is querying within the document, but that was used very sparsely in these sessions.

The first column depicts the sessions of participant P1 from Group 1. The target document in these sessions is shown in D mode. Browsing in these cases starts at beginning of the document (left upper corner). The last two columns illustrate the sessions of participants P2 and P3 from Group 2. In these sessions, the target document is shown in BEP mode and browsing starts at BEP shown by the black line.

In general, the browsing patterns differ by task, representation mode and individual. In all cases the participants do browse around the entry point, that is, they want and need to see more content and context to judge the relevance of the information. Despite variation, some patterns suggested in [5] emerge: step navigation (P1-2), mountain (P2-2) and flatline (P2-3), and all in

Table 2. Number of correct answers per all answers per task

	Task # 1	2	3	4
Group 1	6/8	7/8	2/8	**8/8**
Group 2	4/5	**6/6**	**3/5**	3/4

The results of the tasks in BEP mode are in **bold**. The number of answers varies because all participants did not accomplish all tasks.

combinations. In Task 2, all sessions are connected to correct answers, yet the patterns and times differ.

Task 3 (Kuhlthau) had most variation in browsing patterns and the task was also perceived as the most difficult. Here, the difference between D and BEP is most obvious: less browsing and more dwelling at the entry point in BEP sessions. In Task 3 the entry point is the title of the summary table, thus browsing below the entry point leads to confirming information; yet browsing to the top affords the information that the target document is the right homepage. All sessions but P1-3 are connected to correct answers. In session P1-3 the entry point is visible only for a short while. Session P2-3 is an extreme case: it seems that the participant has identified the relevant information right away and just made a short confirmatory check to the contents of the table.

4.2 A closer look

How to interpret the figures? Like with any log data, the interpretation here is risky. We have paralleled the video taken of the test to the log of the two first sessions of participant P3, and analyze the sessions in more detail. In Task 2, P3 writes altogether nine queries in about 3 minutes, and in between whiles inspects the snippets to improve the queries. He then chooses a document and opens it to a new window (see Figure 1, session P3-2.) None of the queries matched the target document nor its BEP, so without the test manipulation he would not have chosen it. The document is shown in BEP mode. P3 glances at the entry point and reloads the document, perhaps to ensure that he has the right page. Then he starts scrolling down quickly (three steps down in Figure 1, P3-2). From the middle of the document he scrolls swiftly back to entry point and seems to read the document and the task description. After about 10 seconds he scrolls up, to the entry point, and down a bit. He then returns to the entry point and starts writing the answer. While writing, he shortly seems to check the text of the entry point.

In Task 3, P3 writes four queries in two minutes. He inspects the result lists carefully before choosing the first document in the result list of the fourth query. His query matched the target document but not its BEP. He is guided to BEP where he stays for about 1.5 minutes (see Figure 1, P3-3). He looks at the document, then at the task description, then he searches within the document for 'information', and the word is highlighted in the title of BEP. During the first dwell time he mainly reads the task description. The next move he makes is to scroll to the bottom of the document using the space bar. After that follows a 'scrolling-scanning' section in the middle of the session. He swiftly goes to the top and then slowly scans the document back to the entry point. Interestingly, after that he checks twice the top of the document. Then, P3 returns once again to BEP. Now he is obviously convinced about the relevance of the information, and starts to write an answer and perhaps fill in the post-task questionnaire. At the end of the session he routinely scrolls to the top of the document but he does not read it any more.

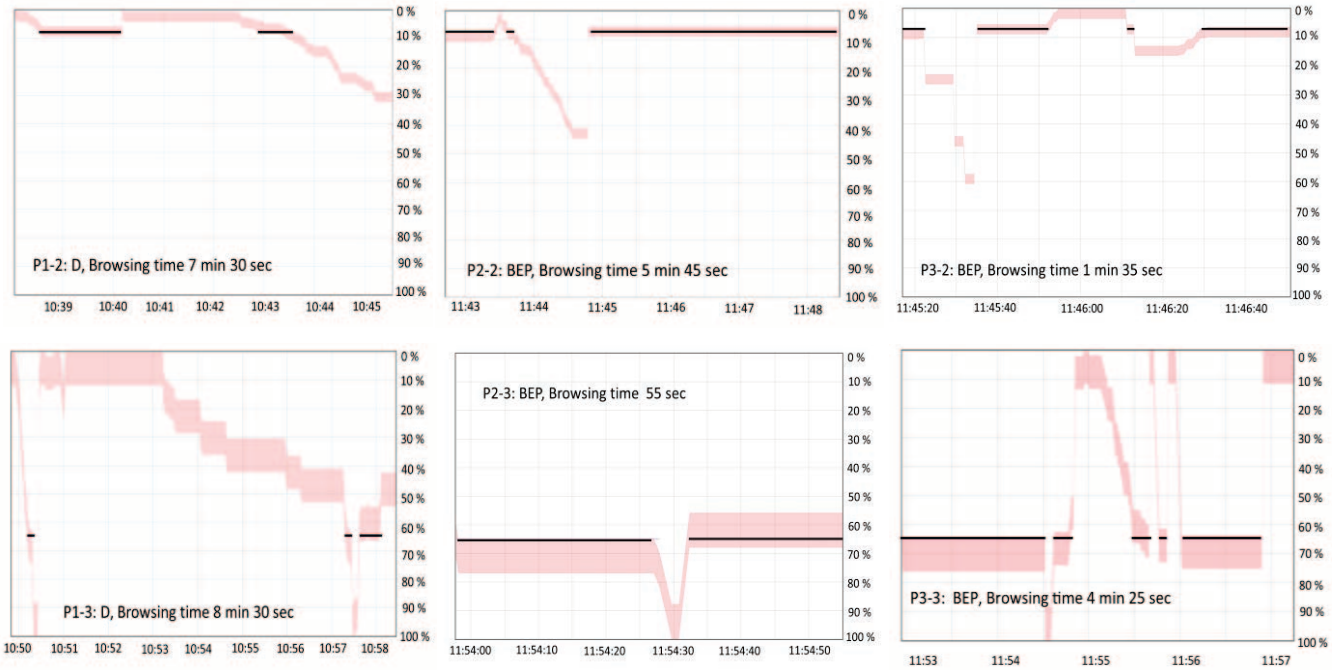

Figure 1. Browsing sessions

5. CONCLUSION

We have explored within document browsing in a lab experiment. Although the test has limitations, like the small number of participants and unbalanced test groups, the results show that document browsing is affected by the search task and its perceived difficulty, the searcher and the document. The best entry point did not seem to give sufficiently content and context for the searchers to judge the relevance. Dwell time may be spent in different activities. Also, there seem to be different strategies for localizing and confirming the relevant information. The browsing patterns found are similar to those suggested by Loizides and Buchanan [5].

The information interaction with retrieved documents has not gained much research focus so far. Although the results of this pilot test are far from conclusive they suggest topics for more profound research: what are the variables, both personal and situational, affecting browsing and reading of retrieved documents? What are the indicators of relevance and how are they recognized?

6. REFERENCES

[1] Lalmas, M., and Reid, J. (2003). Automatic Identification of Best Entry Points for Focused Structured Document Retrieval. In *Proc. of Int. Conf. on Inform. and Knowledge Management.* ACM, New York, NY, USA, 540–543.

[2] Lassila, M., Pääkkönen, T., Arvola, P., Kekäläinen, J., and Junkkari, M. 2012. Unobtrusive Mobile Browsing Behaviour Tracking Tool. In *Proc. of the 4th Information Interaction in Context Symposium*. ACM, New York, NY, USA, 278–281.

[3] Lehtonen, M., Pharo, N., & Trotman, A. 2007. A Taxonomy for XML Retrieval Use Cases. In *Comparative Evaluation of XML Information Retrieval Systems.* LNCS, Vol. 4518, Springer, Heidelberg, 413–422.

[4] Lin, M., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., & Karger, D.R. 2003. What makes a good answer? The role of context in question answering. In *Proc. of the 9th Int. Conf. on Human-Computer Interaction*. s.l., 25-32.

[5] Loizides, L., & Buchanan, G. 2009. An empirical study of user navigation during document triage. In *Research and Advanced Technology for Digital Libraries*. LNCS, Vol. 5714, Springer, Heidelberg, 138-149.

[6] Sigurbjörnsson, B., Kamps, J., and de Rijke, M. 2006. Focused Access to Wikipedia. In *6th Dutch-Belgian Information Retrieval Workshop* (DIR 2006).

[7] Trotman, A., Pharo, N., and Lehtonen, M. 2007. XML-IR Users and Use Cases. In *Comparative Evaluation of XML Information Retrieval Systems*, LNCS, Vol. 4518, Springer, Heidelberg, 400–412

APPENDIX

The search tasks of the test (English translations):

1. Your friend has got a shepherd dog. You think that he should pay a tax for it but he says that shepherd dogs are exempted from taxes. Find out what the law says about this.

2. You have rented a new apartment. Your landlord asks for a rental deposit of three months' rent. You think that this is pretty much. Find out what is a legitimate deposit according to the law.

3. You are writing a thesis about information seeking process model by Carol Kuhlthau. Your supervisor says that there is a good summary of the progression and development of the model on Kuhlthau's homepage. Find the page and give the title of the summary.

4. You have bought a new computer and applied for tax deduction for it. Now you realize that you have lost the receipt although you should have kept it. Find out for how long time the receipts must be preserved for taxation.

[i] light grey in monochrome print