Search and Navigation in Structured Document Retrieval: Comparison of User Behaviour in Search on Document Passages and XML Elements

Gabriella Kazai
Microsoft Research
Cambridge, UK
gabkaz@microsoft.com

Abstract *This paper investigates search and browsing behaviour of users presented with two types of structured document retrieval approaches: passage retrieval and XML element retrieval. Our findings, based on the system logs gathered from 82 participants of the INEX 2006 interactive track experiment (iTrack), indicate that XML element retrieval leads to increased task performance. In addition, qualitative analysis of our video study, where we recorded the interactions of four participants, highlights potential issues with the experimental design employed at iTrack 2006.*

Keywords XML element retrieval, passage retrieval, INEX interactive track, video user study.

1 Introduction

Long before the eXtensible Markup Language (XML) [2] became widely adopted as a standard document format, work in the field of Structured Document Retrieval (SDR) was already underway to address shortcomings of traditional IR. Recognising that users are often only interested in the parts of a document that is relevant to their information need, researchers turned their attention to developing new retrieval approaches to deliver more focused results to users.

The underlying premise of SDR is that the logical structure of a document serves as a source of valuable information that should be exploited in indexing, retrieval and presentation of the document. The ultimate goal is to improve retrieval effectiveness through a more focused retrieval approach, which returns document components to the user, instead of complete documents, thereby reducing users' effort in locating relevant information. The need to consider smaller units within documents as retrievable entities has been considered particularly viable in the case of long documents or documents that cover a variety of topics.

Early examples of SDR include approaches to passage retrieval [3, 7, 9] and hypertext [5]. The more recent revival of SDR is a result of the widespread use of XML on the Web and in digital libraries, which has led to a drastic increase in the number of XML IR systems being developed [1]. Another important catalyst of recent advances in XML IR is that of the INitiative for the Evaluation of XML retrieval (INEX)[1] [6]. Since 2002, INEX has been promoting research in XML IR by providing a forum for researchers to evaluate their XML retrieval approaches and compare their results.

Since 2004, the interactive track (iTrack) at INEX [13, 10] has been investigating the behaviour of users when interacting with components of XML documents. In 2006, iTrack aimed to answer questions regarding the differences and similarities between XML element retrieval and passage retrieval approaches [11].

We took part in the iTrack 2006 experiments with 8 participants. The standard participation involved each participant completing four search tasks using either a passage retrieval or an XML element retrieval system. Data was collected through system logging as well as via questionnaires. In addition, we video recorded four of our participants and administered additional questionnaires. This paper reports our analysis of the log data combined with the qualitative data we collected from the videos. Analysis of the questionnaire data is reported separately in [8].

The paper is structured as follows. Section 2 details the experimental setup. Section 3 presents the results of our analysis of the system logs for 82 users. Section 4 summarises the findings of our video study for our 4 users. Conclusions and future work are reviewed in Section 5.

2 Experimental Setup

The interactive track (iTrack) at INEX 2006 [11] examined users' interaction with XML documents in an experimental laboratory environment. The task focused on comparing XML element retrieval with passage retrieval to explore potential benefits and trade-offs.

2.1 Document Collection and Topics

The experiment used the INEX Wikipedia collection [4] consisting of 659,388 documents totalling about 4.6GB of data and twelve topics (simulated work tasks).

2.2 Search Systems

Participants were asked to perform search using two different search engines: the Panop-

[1] http://inex.is.informatik.uni-duisburg.de/

Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia, December 10, 2007.

ticTM/FunnelbackTM passage retrieval system provided by CSIRO (passage system) and the TopX [12] XML element retrieval system (element system).

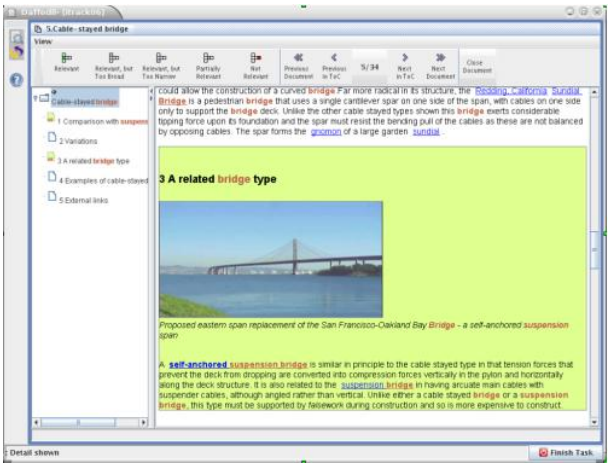

Figure 1: TopX XML element retrieval system.

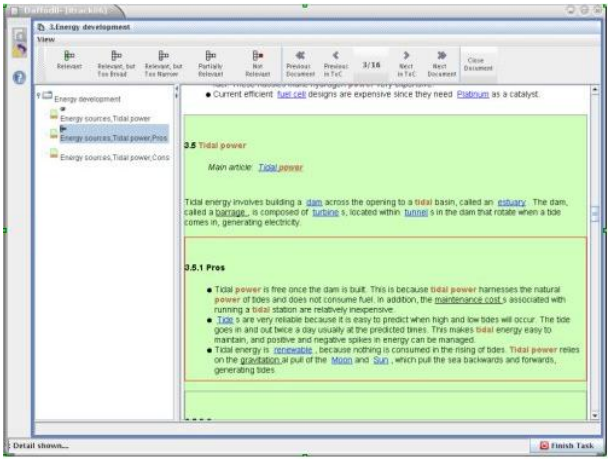

Figure 2: Panoptic passage retrieval system.

In order to remove any experimental bias due to user interface differences, both search engines were interfaced with a consistent look and feel. The similarity of the two systems can be seen in Figures 1 and 2. Figure 1 shows the front end to the TopX search engine, while Figure 2 shows the front end to Panoptic.

The difference between the two systems was subtle. In both, as a response to a user query, the search engine returned an ordered list of documents and, within each document, an ordered list of non-overlapping document parts. The main difference is in the returned retrieval entities: The passage retrieval backend returned non-overlapping passages derived by linearly splitting the documents. The element retrieval system returned XML elements of varying granularity based on the hierarchical document structure. In both versions, the passages and the elements were grouped by the document they belong to. Another important difference is that the table of contents in the document view of the element system included all sections down to a certain level, whereas the table of contents of the passage retrieval system only contained passages that were estimated relevant by the system.

Both systems indicated the parts of the documents that were estimated useful for the searcher in several ways: 1) Up to three high ranking passages/elements were shown per document in the result list. For each, a relevance bar icon indicated the degree of potential usefulness. 2) Relevant document parts were also indicated within the table of contents pane of the document view (see Figures 1 and 2) using the same relevance bar icons. Finally, 3) the relevant document parts were also highlighted in the text of the documents using shades of green and yellow to indicate varying degree of relevance.

2.3 Participants

Seven research teams took part in iTrack 2006, engaging a total of 82 participating users. Our own contribution to this pool was 8 users. Four of our users also participated in a video study, where we recorded their interactions with the search systems to supplement the log data.

We refer to our four participants (three male and one female) as Mark, John, Ben and Eva, respectively. Their average age was 32.25, the youngest being 24, the oldest 45 years old. On average, our users had 11.5 years of search experience on the Web. Summary information can be found in Table 1.

Table 1: Participants of our video study. Search experience is given in years. Acronyms: English (En), Other (Othr), Undergraduate Degree (U.Deg), Usability Consultant (UC), Researcher (Res), Student (Stu).

	Mark	John	Ben	Eva
Gender	M	M	M	F
Age	45	29	24	31
Native lang.	En	En	Othr	Othr
Education	PhD	PhD	U.Deg.	PhD
Occupation	UC	Res	Stu	Res
Web search exp.	10	12	12	12

2.4 Methodology

Initially, participants were presented with a sample topic and given as long as necessary to familiarize themselves with each search system. Each participant then had to complete four search tasks, choosing from a pool of three topics per task (see Table 2). The order in which each category topic was presented to a participant, and the system on which the search task was assigned changed between users as shown in Table 3. Fifteen minutes were allocated to complete a task using one of the search engines.

In addition to the standard track participation, we video recorded the search sessions of four of our participants in order to obtain detailed qualitative information regarding their interaction with the search systems. This allowed us to capture the context of users' interactions and interpret the actions logged by the system. In

Table 2: Topic categorisation.

Category	Topics	Category	Topics
A1	1,2,3	B1	2,3,4
A2	5,6,7	B2	6,7,8
A3	9,10,11	B3	10,11,12
A4	4,8,12	B4	1,5,9

Table 3: Permutations of the topic categories and retrieval systems across participants.

System:		**Element**		**Passage**	
Participant		**Topic categories**			
P1	–	A1	A2	A3	A4
P2	–	A2	A1	A4	A3
P3	–	A3	A4	A1	A2
P4	–	A4	A3	A2	A1
System:		**Passage**		**Element**	
Participant		**Topic categories**			
P5	Mark	B4	B3	B2	B1
P6	John	B3	B4	B1	B2
P7	Ben	B2	B1	B4	B3
P8	Eva	B1	B2	B3	B4

the following section, we describe the findings from the log data analysis for all 82 participants and then discuss in detail the findings from our video recordings.

3 Log Analysis

The collected log data from search tasks of 82 participants comprises 378 logged search sessions, out of which 301 were completed or had valid XML output (some were aborted or restarted). The statistics presented here are based on these 301 session logs.

3.1 Search Sessions

A search session was defined as one experiment involving one participant and one search task. The maximum session duration was hence 15 minutes, the time limit allocated to a task.

Table 4 summarises our findings. From the total of 301 sessions, 173 were search tasks conducted on the passage system and 128 on the element retrieval system. The average session duration was 10 minutes, $2/3$-rd of the allocated time. On average, users spent less time using the passage retrieval system: 9.8 minutes per session, compared with 10.4 minutes using the element retrieval system. This alone, however, does not reveal whether users were successful in completing their tasks within that time. Participants may have given up their search or may have run out of time before finishing a task.

Users averaged 0.49 and 0.74 queries, 0.52 and 0.54 result clicks, and 0.98 and 1.3 visited (browsed) components per minute on the passage and element retrieval systems, respectively. Thus, they issued more queries, clicked on more results in the retrieved list and browsed more inside documents with the element retrieval system than with the passage system. This could be attributed to the differences in search performance, suggesting that the passage system was better at answering users' queries. However, it could also mean that users were better supported in their browsing using the element system. Note that the average query rate per second is only 0.01 across all iTrack experiments, compared with 0.028 reported in [14] for the Web.

3.2 Search Trails

To study users' post-query browsing behaviour, we extracted search trails from the system logs. We defined a search trail as a series of visited document components, initiated with a click on a result in the ranked list, and terminated by a new query or by the end of the session. We extracted 812 trails in total. From these, we calculated the features described in Table 5.

Among 812 trails, 460 were obtained from the 173 logged sessions on the passage system (2.65 trails per session) and 352 from the 128 sessions on the element system (2.75 trails per session). This means that users created roughly the same number of trails in both systems. The average trail length indicates that users visited, on average, 1.4 more document components per query using the element retrieval system (6.9) than with the passage system (5.5). However, they spent on average less time per trail step using the element system (9 seconds vs. 10.3 seconds for the passage system). This suggests that users accessed information in smaller chunks using the element retrieval system, which took less time to skim over but lead to extended trails. A look at the medians on the other hand reveals that the element retrieval system's average trail duration is heavily influenced by outliers.

The logs contained a total of 1580 result clicks (856 in the passage and 724 in the element system). The average rank position of clicks reveals that users looked, on average, further down the ranking with the element system (rank 6 vs. 4.9) than with the passage system.

3.3 Search Success

During the experiments, participants provided relevance feedback by assigning one of five relevance grades (not relevant, partly relevant, too big, too narrow, fully relevant) to some of the visited document components. Table 6 presents statistics on the collected relevance assessments. A total of 2308 judgements were collected: 1169 on the passage (6.8 per session) and 1139 on the element system (8.9 per session). Out of these, 1833 document components were judged relevant: 943 on the passage (5.45 per session) and 890 on the element system (6.95 per session). This means that overall, users found and judged more relevant document components using the element system, thus were likely to have been more successful in completing their task. Comparing this with our reported session durations, we can conclude that although users spent less time on a task using the passage system, they were also less successful. Thus, the reduced session duration

Table 4: Analysis of search sessions for passage and XML element systems.

	Total	Passage	Element
Total sessions (analysed)	301	173	128
Average session duration (mins)	10	9.8	10.4
Average number of queries issued per session	6	4.7	7.7
Average number of queries per minute	0.6	0.49	0.74
Average number of result clicks per session	5.2	4.9	5.7
Average number of result clicks per minute	0.52	0.51	0.54
Average number of browsed components per session	11.2	9.5	13.5
Average number of browsed components per minute	1.12	0.98	1.3

Table 5: Analysis of search trail features for passage and XML element systems. Standard deviation is shown in brackets.

	Total	Passage	Element
Total search trails	812	460	352
Average trail duration (sec)	59 (131.4)	56.9 (140.9)	62.5 (117.8)
Median trail duration (sec)	14	15.3	13
Average trail length	6.1 (7.2)	5.5 (7.1)	6.9 (7.2)
Median trail length	4	4	5
Average step duration (sec)	9.7	10.3	9
Average result click position	5.4 (7.2)	4.9 (6.6)	6 (7.8)
Median result click position	3	2	3

may in fact reflect users' tendency to give up their task using the passage system.

Looking at the average rank of relevant components in the result ranking, however, shows that users of the element system had to look further down the ranking to find relevant information: on average 5.6 ranks vs. only 3.6 ranks for the passage system. This may suggest that while the passage system was better at ranking the results in the ranked list, it was then easier for users to discover more relevant information through browsing using the element system.

4 Video Study

In this sections, we provide a qualitative analysis of our users' search and browsing behaviour from our video recordings with the aim to gain detailed insight into the context of their logged actions.

4.1 Search and Navigation

Our four participants displayed varying strategies to search and navigation in the experiments. We provide a detailed review of their actions based on collected video footage that we cross-referenced with the system logs.

Mark. Mark chose topic 9 (C4) for his first task. Unbeknown to him, he carried out the task using the passage retrieval system. He only issued a single query during the whole session and worked his way down the ranked list in a linear fashion. Even though the task had to be restarted twice due to system problems, he simply re-entered the same query and continued where he left off in the ranking. Initially, his interactions were limited to selecting the whole document as entry point and then scrolling inside the document. His speed of scrolling was influenced by the text-highlights: he would stop or slow down once he reached highlighted parts. Six minutes into the task, he realized that he can navigate inside the document using the table of contents (ToC). From then on, he relied on this form of navigation almost exclusively. He would still click the whole document from the ranked list, but then he would jump straight to a highlighted section using the ToC. He reduced his scrolling activities to the occasional scroll around a highlighted text fragment either for context or in search of further relevant information. In total, Mark viewed 18 document components. The first 7 involved document level entry and scrolling. From the last 11, in 9 instances he entered at the document level, but immediately navigated to a passage using the ToC; and in 2 cases he entered directly by choosing a passage from the ranked list.

Mark's second task was topic 10 (C3), using the passage system. Again, he entered only one query during the whole session, which he repeated when the task was restarted due to a system error. As a result of the system completely crashing then, he only managed to visit 3 results before the task was aborted. In all 3 of these cases, he selected a passage directly from the ranked list and employed no additional scrolling or navigation inside the document.

For his third task, he chose topic 8 (C2) and used the element retrieval system. He issued 9 queries, but only examined results returned for 2 of these. In total, he viewed 6 results: 5 element level and 1 document level entry. He did not do any scrolling inside the document

Table 6: Analysis of search success for passage and XML element systems. Standard deviation is shown in brackets.

	Total	**Passage**	**Element**
Total relevance judgements	2308	1169	1139
Total number of relevant components	1833	943	890
Average rank of judgement	4.9 (7)	4.2 (6.8)	5.7 (8.9)
Median rank of judgement	3	2	3
Average rank of component judged relevant	4.6 (6.1)	3.6 (5.5)	5.6 (8)
Median rank of component judged relevant	2	2	3

in 4 out of the 6 instances and only once clicked on an entry in the ToC.

In his final task, he worked on topic 3 (C1) using the element system. This was the only task that he successfully completed (as stated by himself in the post-task questionnaire). He issued 2 queries and viewed 3 results in total: 2 element-level and 1 document level. All 3 cases involved extensive scrolling inside the document and no interactions with the ToC.

In summary, Mark's interactions show a learning curve and point towards preference for the SDR approach, i.e. having direct access to relevant text fragments. His initial strategy for navigating inside a document using the ToC was later replaced by extensive scrolling. Once he became confident with selecting passage level entries, he completely stopped using the ToC for navigation. When scrolling, he would typically stop at the first highlighted fragment. He would also usually scroll around a relevant fragment to obtain more context.

John. In his first task (topic 11 in C3 on the passage system), John issued 2 queries and viewed 3 results. In all three cases, he chose the document as his entry point but then he made extensive use of the ToC to navigate (clicked a total of 9 ToC links). He also scrolled frequently in both directions (up and down) from an entry point. Due to UI design issues, whereby scrolling in the right document view pane was independent from the item selection in the ToC pane, he got disoriented on multiple occasions. To reorient himself, on one occasion he browsed through the ToC (adding an additional 10 ToC clicks to the log, which reported a total of 19 ToC clicks).

He initiated 5 queries in his second task (topic 9 in C4 on the passage system) and viewed 7 results. For one query, he scrolled all the way till the end of the ranking but did not click any results. For all 7 clicked results he chose document level entry points and scrolled inside the document.

For his third task (topic 3 in C1 on the element system), he ran 7 queries and viewed 6 results, all via document level entry. Whilst viewing the fourth document, the system developed a fault with the scrollbar. In an effort to try to get around this, he clocked up an additional 9 clicks on various ToC links.

For his final task (topic 6 in C2 on the element system), he ran 3 queries and viewed only one returned result for each. The only time he clicked a passage level entry was for the final result, but then he scrolled to the top of the document and back down again. For 2 of the 3 viewed documents, he relied exclusively on scrolling to navigate inside the document. This was the only task he failed to complete.

In summary, John followed his own established search and navigation strategy. He was familiar with standard retrieval systems: he chose document level entry points and scrolling for navigation. Unlike Mark, John's search and navigation strategy was constant throughout. He used many queries but would, on average, view one result per query. Unless a relevant document was presented within the top 2 ranks, he would quite often scroll through the whole ranking before making his selection. He would always choose the whole document as entry point. Inside the document, he would always first browse by scrolling and then combine the use of ToC with further scrolling. The speed of his scrolling was influenced by the text highlights: he would slow down when scrolling over highlighted document parts. He would also scroll around highlighted areas for context.

Ben In his first task (topic 7 in C2 on the passage system), he ran 5 queries and viewed 1 document for each. In each case, he chose passage level entry points into the documents. He kept scrolling to a minimum and only once used the ToC for navigation.

For his second task (topic 2 in C1 on the passage system), he issued 4 queries, 1 of which did not result in any viewed documents. From the total of 6 viewed documents, 3 were reached through document level entry points. For navigating inside the documents, he employed the use of both the ToC (he clicked 6 ToC entries in total) and scrolling. He regularly scrolled up in a document to revisit already skim-read sections.

During his third task (topic 1 in C4 on the element system), he entered 7 queries, 2 of which were unsuccessful in providing any documents of interest to Ben. He viewed 7 documents, 3 of which he entered at the document level. For this task, he only once clicked on a link in the ToC. The rest of the time he simply scrolled. The text highlighting did not influence his speed of scroll.

His final task (topic 10 in C3 on the element system) showed a different behaviour. For 9 issued queries he viewed 8 documents, 7 of which by document level

entry. He regularly clicked around in the ToC or scroll around in a document, but did not combine both forms of navigation inside a single document.

In summary, Ben's search and navigation behaviour was a mixed bag. He combined all methods and was as successful in completing his tasks as John (only failed to finish task 4). His strategy was very flexible, easily adapting to the task at hand. He confidently used the various forms of navigational methods inside a document. He varyingly chose between document level entry points and directly accessing relevant text fragments from the ranking.

Eva For her first task (topic 3 in C1 on the passage system), Eva ran 11 queries, but viewed only 4 results. After inputting a query, she would inspect the ranked list scrolling down till rank 9, on average. Once she found a relevant document, she spent on average 3.6 minutes on browsing and reading through it. She always chose document level entry points. Inside a document she would typically scroll. Although she also generated 16 ToC hits, in her own words, these were only clicked in order to allow her to make relevance judgments.

For her second task (topic 8 in C2 on the passage system), she used 5 queries, but inspected only 3 results due to system errors. For all 3 results she chose passage level entry points. She only used the ToC on one occasion in this task, which was again related to system error: She clicked a passage level entry point, but found that the document part shown in the right pane did not match the ToC entry. In order to realign the screens she toggled between entries in the ToC.

In her third task (topic 10 in C3 on the element system), she ran 4 queries and viewed 5 documents, all through passage level entries. Again she only used the ToC in order to provide relevance feedback. She completed this task just under 10 minutes.

For her final task (topic 1 in C4 on the element system), she issued 15 queries, but failed to complete the task. After carefully looking through the result lists, she only chose to view 7 documents, accessing them directly at the passage level. Inside the document, she scrolled relatively little and again only clicked in the ToC before making a relevance judgment.

In summary, Eva tended to either only choose document level or passage level entries, but not both types within an individual task. Although she stated that she only selected items in the ToC to make judgments, in her first task, she actually did make quite a lot of use of the ToC. She has even adjusted the UI to show more of the ToC and she spent long periods of time looking through and reading the ToC. Eva managed to complete half of her tasks (tasks 2 and 3).

4.2 User Opinions on Passage vs. Element Retrieval

In a post-task questionnaire, users rated their search experiences after each task using a 5 degree scale (1=frustrating, 3=neutral, 5=pleasing). Table 7 shows the averaged ratings for the two systems. Based on these numbers it appears that the XML element retrieval system is preferred by our participants as it leads to higher user satisfaction. In order to find out what criteria these overall ratings were formed, we asked participants to explain their reasoning for the assigned score.

Mark told us that he was frustrated with the passage retrieval system as it crashed for his first two tasks. He was hence unable to find any relevant information to complete the tasks. He was happier with the element retrieval system as it did actually work and he also managed to solve his fourth task. His rating in this case was based on the fact that the system "performed better" and that he "was able to find all the necessary information" for his task.

John rated both systems equal on average. His explanation was that "compared to Google or Wikipedia, the task took [him] about the same time [to complete using these systems]. Usually [he] has more windows open but the current user interface was not very flexible." He also commented that a lot of the new user interface features, such as the ToC, he did not use as he was more familiar with Web search. So, his rating was based on a combination of system performance (effectiveness and speed) and user interface features. His rating was not influenced by his success or failure in completing a task (e.g. he gave a rating of 2 after his successful completion of task 3, and a rating of 4 after he failed task 4).

Ben again rated the two systems equal on average. His rating was a reflection on a mixture of criteria: whether he was successful in his task, the retrieval effectiveness of the system (whether it was able to return relevant hits: "nothing directly mentioned peaceful revolution"), system features and their usefulness. For example he commented "the related terms was useful here", "the table of contents was quite hard to use here", and "the extra features didn't help here".

Eva's ratings were mostly dependent on her completion of the task. Additional factors for her task 1 rating included some user interface aspects: "I am used to how I search on the Web. [As] this task has sub-tasks, I would start several browsers for each sub-task. I could not do this here."

From the range of replies, we can see that a system's ability to locate relevant information is one of the main factors of user satisfaction. This is then supplemented by factors such as efficiency and general user experience supported by system features and user interface.

In a post-experiment questionnaire, users were asked to identify the differences between the two systems. Our participants' answer to this was unanimous: None of them were able to identify any difference between the element and passage retrieval systems. This would suggest that the use of structure (as exposed by the two systems) did not play a role

in our users satisfaction with the systems. However, once the differences were explained to them, they did comment that the ToC for the element retrieval system was more useful as it allowed navigation to any structural part of a document. The passage retrieval system at best had three passages listed in its ToC. This has, in fact, presented an issue to John, who was thus unable to provide judgement for a passage he thought relevant, but which was not listed in the ToC. Since he had no means of adding the passage to the ToC, he was unable to provide relevance judgement for it.

Table 7: Averaged user rating of search experience for passage retrieval system (P) and XML element retrieval system (E).

	Passage	**Element**
Mark	1	3.5
John	3	3
Ben	3.5	3.5
Eva	3	3.5
Overall	2.6	3.4

4.3 System Logs vs. Video Evidence

During the combined analysis of the system log and the video study data, we came across a number of anomalies. This has lead us to uncover some issues with the iTrack experiment where the log data may lead to incorrect conclusions. For example, the logs reported a very high ratio of within-document navigation using the ToC (over 22% of all logged actions over all 82 participants; with average trail length of 6.1 steps). This would suggest to system designers that the table of contents was a very useful navigation tool for users. However, our video recordings indicate that this is likely to be a highly over-exaggerated figure. We found that users may click repeatedly on entries in the ToC in response to system hickups and errors, or as workarounds for system design faults. For example, both Ben and John used the ToC links (and clicked back and forth several times) simply to re-orient themselves within a document after the display crashed. Both Mark and Eva had to close documents and then re-open them due to the system's failure to display the document's content. They would then often click several links in the ToC just to check if the system was still responding. In addition, John clicked on ToC entries just to re-align the two panes of the display after he has been scrolling up and down in a document.

Another reason why participants clicked links in the ToC was to make relevance judgments. This was a system-imposed limitation whereby only document parts selected in the ToC could be judged and only when users selected the document part using the ToC. This has lead to a large increase in ToC navigation clicks. Based on our video records that we cross-referenced with the system logs we estimate that for our 4 participants less than half of the logged ToC clicks were valid navigational user actions. Over 50% is attributed to consequences of system design constraints or system errors and crashes.

Additional issues with the systems included that some ToC links were incorrect, their title text was wrong or they took the user to the incorrect location. This has again led to increased user navigation, but also resulted in users more frequently abandoning a particular search inside a document.

We have also witnessed users mistakenly assigning a relevance score to the wrong document part. John, for example, has on numerous occasions lost the synchronization between the left and right panes of the document view as he often scrolled inside the document. As a consequence, he has on two occasions marked the wrong section of the document relevant (i.e., the selected part in the ToC was different to what was actually shown on the screen). The same problem showed up in all our participants' videos.

5 Conclusions

In this paper we reported on our experience of participating in the INEX 2006 iTrack experiments. We provided an analysis of the system logs collected for all 82 participants at iTrack.

In addition, we gave a detailed account of four of our participants' search and navigation behaviour based on combined evidence extracted from the logs and from our video study. We found that users have their own personal styles of searching and navigating. Some users adopt new strategies easily while others prefer to stick with tried and tested methods. An obvious implication for the design of SDR systems is that any new forms of interaction and navigation has to be supported by simple and self-explanatory user interface features. Furthermore, the interaction model needs to be useful enough to promote its use.

When comparing passage and XML element retrieval methods, we found evidence to suggest that element retrieval led to increased task performance with more document components found and judged relevant. This was achieved by users at a cost of spending on average more time on a task and issuing more queries per session. On average, users would also browse more with an element retrieval system, leading to an average search trails of 6.9 visited document components (vs. 5.5 using the passage system). Both are, however, above those reported in [14] for Web search.

Furthermore, our users rated on overall the element retrieval system above the passage system (without knowing that they were rating two different systems). We also found evidence to suggest that users' navigation behaviour differs across the two systems: Participants were more likely to select full documents as entry points using the passage system and then make more extensive use of the ToC. In element retrieval, participants more often chose document parts as entry points but were then less likely to use the ToC for

navigation. This can be motivated by the argument that once users gained direct access to relevant parts, they did not need to navigate as much inside the document.

Finally, our investigation has highlighted a possible issue with the experimental design adopted at iTrack. We found evidence that users' behaviour was inadvertently affected by a system imposed constraint: In order to provide relevance judgements on a document component, users had to ensure that the component was selected in the ToC. This meant that users were often forced to click an entry in the ToC before being able to make relevance judgements, and thus increasing ToC click statistics in the logs. One consequence of this is that the average search trail length reported here could well be an overestimate.

A limitation of this study lies in the question of the generality of its findings. While the set of 82 users of the overall iTrack experiments represent a sufficiently large user population, experienced system errors and crashes raise questions on the fidelity of the collected data. For example, out of the total 378 sessions 59 were generated as a result of 41 restarted search tasks (some of which were restarted multiple times) affecting 33 users in total. On the other hand, our video study only included a small group of 4 participants.

Another issue concerns the Wikipedia collection used in the experiments, where most articles are short or are divided into small chunks. Such a collection may not be best served by passage retrieval techniques as highlighted in [7]. Further studies are thus needed to determine if users prefer passage or element retrieval systems.

Our future work will extend the analysis presented here to compare systems on a per topic basis. We will also look to run user studies on a larger scale incorporating system logging and video capture. Our aim is to build on our findings in order to design appropriate user interfaces for SDR. In addition, we hope that our conclusions regarding the analysis of the system logs will help other iTrack 2006 participants in their work.

Acknowledgements Many thanks to Natasa Milic-Frayling, Birger Larsen and the anonymous reviewers for their many helpful comments.

References

[1] Henk M. Blanken, Torsten Grabs, Hans-Jörg Schek, Ralf Schenkel and Gerhard Weikum (editors). *Intelligent Search on XML Data, Applications, Languages, Models, Implementations, and Benchmarks*, Volume 2818 of *LNCS*. Springer, 2003.

[2] T Bray, J Paoli and C.M. Sperberg-McQueen. Extensible Markup Language (XML) 1.0. http://www.w3.org/TR/1998/REC-xml-19980210, W3C Recommendation. Technical report, W3C (World Wide Web Consortium), February 1998.

[3] James P. Callan. Passage-level evidence in document retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310, New York, USA, 1994. Springer-Verlag New York, Inc.

[4] Ludovic Denoyer and Patrick Gallinari. The wikipedia xml corpus. In Fuhr et al. [6], pages 12–19.

[5] Mark Edwin Frisse. Searching for information in a hypertext medical handbook. In *HYPERTEXT '87: Proceeding of the ACM conference on Hypertext*, pages 57–66, New York, NY, USA, 1987. ACM Press.

[6] Norbert Fuhr, Mounia Lalmas and Andrew Trotman (editors). *Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Dagstuhl Castle, Germany, December 17-20, 2006, Revised and Selected Papers*, Volume 4518 of *Lecture Notes in Computer Science*. Springer, 2007.

[7] Marcin Kaszkiel and Justin Zobel. Effective ranking with arbitrary passages. *JASIST*, Volume 52, Number 4, pages 344–364, 2001.

[8] Gabriella Kazai and Andrew Trotman. Users' perspectives on the usefulness of structure for XML information retrieval. In *Proceedings of the 1st International Conference on the Theory of Inofrmation Retrieval*, pages 247–260, 2007.

[9] Xiaoyong Liu and W. Bruce Croft. Passage retrieval based on language models. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 375–382, New York, NY, USA, 2002. ACM Press.

[10] Saadia Malik, Birger Larsen and Anastasios Tombros. Report on the INEX 2005 interactive track. *SIGIR Forum*, Volume 41, Number 1, pages 67–74, 2007.

[11] Saadia Malik, Anastasios Tombros and Birger Larsen. The interactive track at INEX 2006. In Fuhr et al. [6], pages 387–399.

[12] Martin Theobald, Ralf Schenkel and Gerhard Weikum. An efficient and versatile query engine for TopX search. In Klemens Böhm, Christian S. Jensen, Laura M. Haas, Martin L. Kersten, Per-Åke Larson and Beng Chin Ooi (editors), *VLDB*, pages 625–636. ACM, 2005.

[13] Anastasios Tombros, Saadia Malik and Birger Larsen. Report on the INEX 2004 interactive track. *SIGIR Forum*, Volume 39, Number 1, pages 43–49, 2005.

[14] Ryen W. White and Dan Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 255–262, New York, NY, USA, 2007. ACM Press.