The Interactive Track at INEX 2004

Anastasios Tombros[1], Birger Larsen[2], and Saadia Malik[3]

[1] Dept. of Computer Science, Queen Mary University of London, London, UK
tassos@dcs.qmul.ac.uk
[2] Dept. of Information Studies, Royal School of LIS, Copenhagen, Denmark
blar@db.dk
[3] Fak. 5/IIS, Information Systems, University of Duisburg-Essen, Duisburg, Germany
malik@is.informatik.uni-duisburg.de

Abstract. An interactive track was included in INEX for the first time in 2004. The main aim of the track was to study the behaviour of searchers when interacting with components of XML documents. In this paper, we describe the motivation and aims of the track in detail, we outline the methodology and we present some initial findings from the analysis of the results.

1 Interactive Track Motivation

In recent years there has been a growing realisation in the IR community that the interaction of searchers with information is an indispensable component of the IR process. As a result, issues relating to interactive IR have been extensively investigated in the last decade. A major advance in research has been made by co-ordinated efforts in the interactive track at TREC. These efforts have been in the context of unstructured documents (e.g. news articles) or in the context of the loosely-defined structure encountered in web pages. XML documents, on the other hand, define a different context, by offering the possibility of navigating within the structure of a single document, or of following links to another document.

Relatively little research has been carried out to study user interaction with IR systems that take advantage of the additional features offered by XML documents, and so little is known about how users behave in the context of such IR systems. One exception is the work done by [1] , who studied end user interaction with a small test collection of Shakespeare's plays formatted in XML.

The investigation of the different context that is defined in the case of user interaction with XML documents has provided the main motivation for the establishment of an interactive track at INEX. The aims for the interactive track are twofold. First, to investigate the behaviour of users when interacting with components of XML documents, and secondly to investigate and develop approaches for XML retrieval which are effective in user-based environments.

In the first year, we focused on investigating the behaviour of searchers when presented with components of XML documents that have a high probability of being relevant (as estimated by an XML-based IR system). Presently, metrics that are used for the evaluation of system effectiveness in the INEX ad-hoc track are based on certain

N. Fuhr et al. (Eds.): INEX 2004, LNCS 3493, pp. 410–423, 2005.

assumptions of user behaviour [2]. These metrics attempt to quantify the effectiveness of IR systems at pointing searchers to relevant components of documents. Some of the assumptions behind the metrics include that users would browse through retrieved elements in a linear order, that they would "jump" with a given probability p from one element to another within the same document's structure, that they would not make use of links to another document, etc. These assumptions have not been formally investigated in the context of XML retrieval; their investigation formed the primary aim for the first year of the interactive track.

Since the investigation of user behaviour forms our primary focus, the format of the track for the first year differs to that typically followed by, for example, the interactive track at TREC. The main difference was that a comparison between different interactive approaches was not our main focus. Instead, a more collaborative effort was planned, with the outcome of the studies expected to feed back to the INEX initiative. Participating sites still had the option to develop and evaluate their own interactive approaches, but this was not a requirement for participation. It should be noted that none of the participating sites opted to develop their own system.

We first describe the experimental setup and methodology in section 2, then we present an initial analysis of the data in section 3, and we conclude in section 4.

2 Experimental Setup

In this section we outline the experimental set up for the first interactive track at INEX.

2.1 Topics

We used content only (CO) topics from the INEX 2004 collection. We added an additional dimension to the investigation of this year's interactive track by selecting topics that corresponded to different types of tasks. The effect that the context determined by task type has on the behaviour of online searchers has been demonstrated in a number of studies e.g. [3].

One way to categorise tasks is according to the "type" of information need they correspond to. In [3] the categorisation included background (find as much general information on a topic as possible), decision (make a decision based on the information found) and many-items task (compile a list of items related to the information need) types. It was shown that different task types promote the use of different criteria when assessing the relevance of web pages. It is likely that a similar effect, in terms of user behaviour within structured documents, may exist in the context of XML documents. Searchers may exhibit different browsing patterns and different navigational strategies for different task types.

Four of the 2004 CO topics were used in the study, and they were divided into two task categories:

- Background category (B): Most of the INEX topics fall in this category. The topics express an information need in the form of "I'd like to find out about X". The two tasks in this category were based on topics 180 and 192.

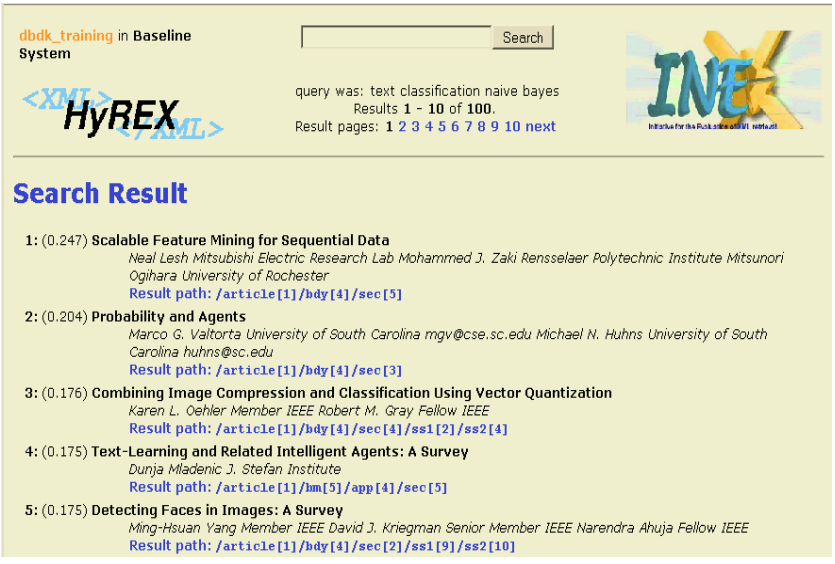

Fig. 1. The ranked list of documents in the Baseline system

– Comparison category (C): There are a number of topics whose subject is along the lines of: ”Find differences between X and Y”. The tasks given in this category were based on topics 188 and 198.

In order to make the tasks comprehensible by other than the topic author, it was required that all INEX 2004 topics not only detail *what* is being sought for, but also *why* this is wanted, and in what context the information need has arisen. Thereby the INEX topics are in effect simulated work task situations as developed by Borlund [4, 5]. Compared to the regular topics, more context on the motives and background of the topic is provided in the simulated work tasks. In this way, the test persons can better place themselves in a situation where they would be motivated to search for information related to the work tasks. The aim is to enable the test persons to formulate and reformulate their own queries as realistically as possible in the interaction with the IR system. The task descriptions used in the study were derived from part of the Narrative field. We include the task descriptions as given to searchers in the Appendix.

2.2 System

A system for the interactive track study was provided by the track organisers. The system was based on the HyREX[1] retrieval engine, and included a web-based interface with a basic functionality.

Searchers were able to input queries to the system. In response to the query, HyRex returns a ranked list of components as shown in Figure 1. The information presented

[1] http://www.is.informatik.uni-duisburg.de/projects/hyrex/

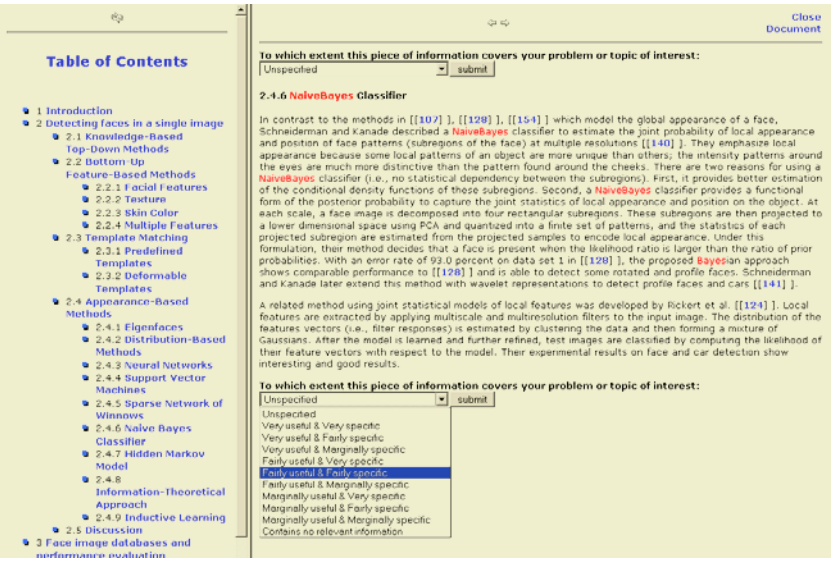

Fig. 2. Detailed view of document components in the Baseline system

for each retrieved component included the title and authors of the document in which the component occurs, the component's retrieval value and the XPath of the component. Searchers can explore the ranked list of components, and can visit components by clicking on the Xpath in the ranked list.

In Figure 2 we show the detailed component view. This view is divided into two parts: the right hand of the view includes the actual textual contents of the selected component; the left side contains the table of contents for the document containing the component. Searchers can access other components within the same document either by using the table of contents on the left, or by using the next and previous buttons at the top of the right part of the view.

Table 1. The applied relevance scale

A	Very useful & Very specific
B	Very useful & Fairly specific
C	Very useful & Marginally specific
D	Fairly useful & Very specific
E	Fairly useful & Fairly specific
F	Fairly useful & Marginally specific
G	Marginally useful & Marginally specific
H	Marginally useful & Marginally specific
I	Marginally useful & Marginally specific
J	Contains no relevant information
U	Unspecified

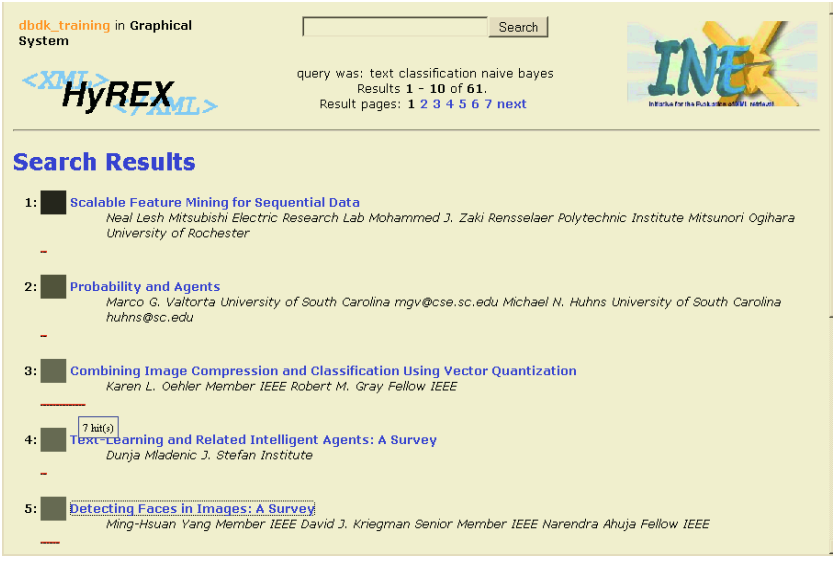

Fig. 3. The ranked list of documents in the Graphical system

A relevance assessment for each viewed component could be given, as shown in Figure 2. The assessment was based on two dimensions of relevance: how useful and how specific the component was in relation to the search task. The definition of usefulness was formulated very much like the one for Exhaustivity in the Ad hoc track, but was labelled usefulness, which might be easier for users to comprehend. Each dimension had three grades of relevance as this is shown in Figure 2 . Ten possible combinations of these dimensions could be made.

To return to the ranked list, searchers would need to close the currently open document. A different version of the system with graphical features was also developed. This system (Graphical system) differed to the Baseline system both in the way of presenting the ranked list (Figure 3) and in the way of presenting the detailed view of components (Figure 4). The graphical system retrieves documents rather than components, and presents the title and authors of each retrieved document. In addition, it also presents a shaded rectangle (the darker the colour the more relevant the document to the query) and a red bar (the longer the bar the more query hits are contained in the document).

The detailed view for each selected document component is similar to that for the Baseline system, with the addition of a graphical representation at the top of the view (Figure 4). A document is represented in a rectangular area and is split horizontally and vertically to represent the different document levels. Tooltips (on mouse-over) provide additional information about the retrieved components, such as the first 150 characters of the contents and the component's name, the selected section, subsection, etc. On the top part of the this view, all the retrieved documents are shown as small rectangles in gray shades along with the Next and Previous links to allow navigation between the retrieved results.

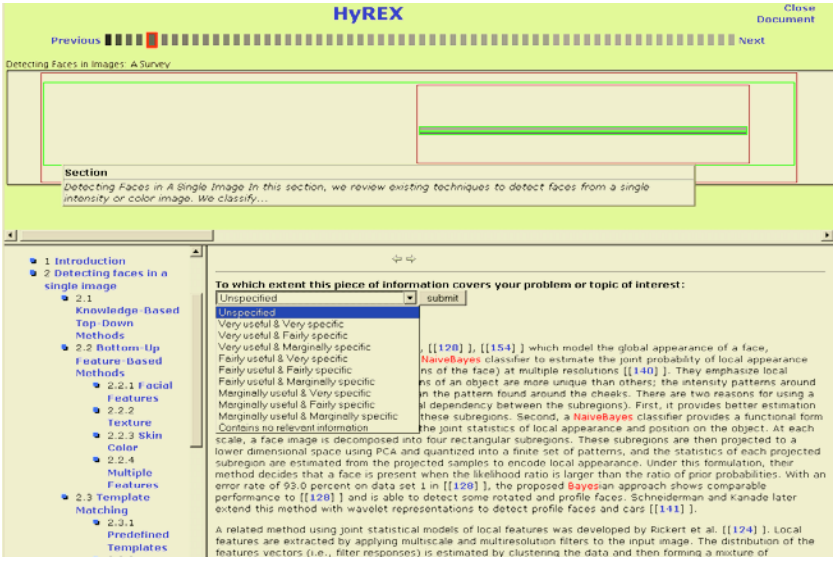

Fig. 4. Detailed view of document components in the Graphical system

2.3 Participating Sites

The minimum requirement for sites to participate in this year's interactive track was to provide runs using 8 searchers on the Baseline version of the XML retrieval system that the track organisers provided. In addition to the minimum requirement, sites could choose to employ more users, to expand the experimental design by comparing both versions of the system (baseline and graphical), or to test their own experimental system against the baseline system provided.

Ten sites participated in the Interactive track. We give the sites' name, number of searchers used and types of comparisons performed in Table 2.

Table 2. Participating sites in the Interactive Track

Site	Baseline System	Additional studies
Oslo University College, Norway	8 users	-
RMIT Australia	16 users	-
U. Twente/CWI, The Netherlands	8 users	8 users(baseline vs. graphical)
Norwegian University of Science and Technology	8 usres	-
U. Tampere, Finland	8 users	-
Kyunpook National University, Korea	8 users	-
Robert Gordon University, Scotland	8 users	-
University of Duisburg-Essen, Germany	8 users	-
Royal School of LIS, Denmark	8 users	-
Queen Marry University of London, England	8 users	-

2.4 Experimental Protocol

A minimum of 8 searchers from each participating site were used. Each searcher searched on one task from each task category. The task was chosen by the searcher. The order in which task categories are performed by searchers was permuted. This means that one complete round of the experiment requires only 2 searchers. The minimum experimental matrix consisted of the 2x2 block shown in Table 3.

This block was repeated 4 times for the minimum requirements for participation. This matrix could be augmented by adding blocks of 4 users (a total of 12, 16, 20, etc. users).

For the comparison of the baseline and the graphical systems, searchers would be involved in the study in addition to the ones used only for the baseline system. The experimental matrix in this case consisted of the blocks of system-task conditions given in Table 4. The order of an experimental session was as follows:

1. Introduction: Briefing about the experiment and procedures
2. Before-experiment questionnaire
3. Hand out Instructions for Searchers
4. System tutorial
5. Task selection from the appropriate category
6. Before-task questionnaire
7. Search session
8. After-task questionnaire
9. Repeat steps 5-8 for the other task category
10. After-experiment questionnaire
11. Informal discussion/interview: any additional views on the experiment, system, etc. the searcher wishes to share.

Each searcher was given a maximum of 30 minutes to complete each task. The goal for each searcher was to locate sufficient information towards completing a task.

2.5 Data Collection

The collected data comprised questionnaires completed by the test persons, the logs of searcher interaction with the system, the notes experimenters kept during the sessions and the informal feedback provided by searchers at the end of the sessions.

The logged data consisted of the queries issued, the components returned by the system, the components actually viewed and the order in which they were viewed, relevance assessments of these, any browsing behaviour, as well as time stamps for each interaction between searchers and the system.

Table 3. Basic experimental matrix

Searcher	**1^{st} Task Category**	**2^{nd} Task Category**
1	Background(B)	Comparison(C)
2	Comparison(C)	Background(B)

Table 4. Augmented experimental matrix

Searcher	1^{st} Condition	2^{nd} Second Condition
1	Graphical-B	Baseline-C
2	Graphical-C	Baseline-B
3	Baseline-B	Graphical-C
4	Baseline-C	Graphical-B

3 Initial Results Analysis

In this section we present an initial analysis of the collected data. In section 3.1 we analyse data collected from the questionnaires, then in section 3.2 we present some general statistics collected from the system logs, and in section 3.3 we outline the detailed analysis of browsing behaviour which is currently in progress.

3.1 Questionnaire Data

A total of 88 searchers were employed by participating sites. The average age of the searchers was 29 years. Their average experience in bibliographic searching in online digital libraries, computerised library catalogs, WWW search engines etc. was 4, on a scale from 1 to 5 with 5 signifying highest experience level. The education level of the participants spanned undergraduate (39%), MSc (49%), and PhD (12%) levels.

In terms of task selection, from the Background task category 66% of participants selected task B1 (cybersickness, topic 192) and 34 % selected B2 (ebooks, topic 180). From the Comparison task category, 76% selected task C2 (Java-Python, topic 198) and 24% selected task C1 (Fortran90-Fortran, topic 188).

In Table 5 we present data for task familiarity, task difficulty and perceived task satisfaction. With respect to task familiarity, we asked searchers before the start of each search session to rate how familiar they were with the task they selected on a scale from 1 to 5, with 5 signifying the greatest familiarity. With respect to task difficulty, we asked searchers to rate the difficulty of the task once before the start of the search session, and once the session was completed (pre- and post- task difficulty, columns 3 and 4 respectively). Searchers also indicated their satisfaction with the results of the task. All data in Table 5 correspond to the same 5-point scale.

The data in Table 5 suggest that there are some significant differences in the searchers' perceptions of the tasks. The most notable of these differences are in task

Table 5. Searchers' perceptions of tasks

	Task familiarity	Pre-task difficulty	Post-task difficulty	Task satisfaction
B1 (no.192)	2.1	2.03	1.47	3.39
B2 (no.180)	2.73	2.1	1.97	1.97
C1 (no.188)	2.67	1.95	1.74	2.62
C2 (no.198)	2.91	2.1	1.52	2.9

Table 6. Access modes to viewed components

Access	B	C	Total	B	C
nextprev	17	17	34	2%	2%
rankedlist	588	550	1138	63%	62%
structure	327	327	654	35%	37%
Total	932	894	1826	100%	100%

familiarity and task satisfaction. It should be noted that at this time a thorough statistical analysis of the results has not been performed. An initial analysis of the correlation between task familiarity and satisfaction did not show a strong relationship between these two variables across the tasks.

The overall opinion of the participants about the Baseline system was recorded in the final questionnaire they filled in after the completion of both tasks. Participants generally felt at ease with the system, finding it easy to learn how to use (average rating 4.17), easy to use (3.95) and easy to understand (3.94). There were also many informal comments by the participants about specific aspects of the system. These comments were recorded by the experimenters and will be analysed at a later stage.

3.2 General Statistics

This analysis concerns approximately 50% of the log data for the baseline system. The remainder could not be analysed reliably at present because of problems with the logging software.

Ranks. A maximum of 100 hits were presented to searchers on the ranked list, and they were free to choose between these in any order they liked (See Figure 1). For the Background (B) tasks 86% of the viewed components were from top10 of the ranked list (80% for the Comparison (C) tasks). The ranks viewed furthest down the list were 71 for B and 96 for C.

Queries. The possible query operators were '+' for emphasis, '-' for negative emphasis, and " " for phrases. The phrase operator was used 24 times in B, and 16 in C. No one used plus or minus. 217 unique queries were given for B, and 225 for C across all searchers. On average, the queries for B consisted of 3.0 search keys (counting a phrase as one search key), and 3.4 for C including stop words. 81% of the queries for B consisted of 2, 3 or 4 search keys for B, 80% for C.

Viewed Components. In total, searchers viewed 804 different components for B, and 820 for C. On average this was 10.9 unique components viewed for B, and 10.8 for C.

Three possibilities existed for accessing a component: to click a hit from the ranked list, to click a part of the document structure (via the table of contents), and to use the next/previous buttons. From Table 6 below it can be seen that very few chose to use the next/previous buttons: only 2% of component viewing arose from this (both B and C). For B 63% of viewings came from the ranked list, for C this was 62%. For B 35% came from the table of contents, and 37% for C.

Assessed Components. 503 components were assessed for B, 489 for C, or 6.8 per searcher per task for B, and 6.4 for C. This corresponds to 63% of the viewed components for B and 60% for C. In 8 cases the searchers pressed 'Submit' without selecting a relevance value (recorded as U in Tables 1, 7, 8 and 9).

The distribution of relevance assessments on tasks can be seen in Table 7 below. It may be observed that 12-13% of the assessed documents were 'Very useful & Very specific" [A] for both B and C, and that 15-16% of the assessed documents were 'Marginally useful & Marginally specific" [I] for both B and C. The most noteworthy difference is that B had 38% non-relevant assessments [J], and C only 17%.

Table 7. Relevance assessments distributed on task type (see Table1 above for relevance scale)

Relevance	B	C	Total	B	C
A	65	61	126	13%	12%
B	28	36	64	6%	7%
C	8	13	21	2%	3%
D	19	45	64	4%	9%
E	36	61	97	7%	12%
F	28	38	66	6%	8%
G	12	20	32	2%	4%
H	33	47	80	7%	10%
I	79	80	159	16%	16%
J	191	84	275	38%	17%
U	4	4	8	1%	1%
Total	503	489	992	100%	100%

The next two tables show the distribution of relevance assessments on the access possibilities, one for B and one for C (i.e. how did the searchers reach the components which they assessed). The total number of component viewings with relevance assessments is lower (992) than the total number of components viewed (1826, Table 6) because not all viewed components were assessed.

For both B and C very few viewings with next/previous section buttons resulted in assessments: 0 for C, and 5 for B. The latter 5 were given low assessments. In both cases the majority of assessments resulted as a direct consequence of clicking a hit from the ranked list: 67% for B and 71% for C. Apart from 1% next/previous navigation in B the remainder the rest is taken up by navigation from the table of contents. Large variations are, however, obvious in the data, and can be uncovered by an in-depth analysis of the browsing behaviour.

Overall Browsing Behaviour. Table 10 shows this variation on an overall level by counting the number of requests for components within the same document. The raw figures included double counting, because whenever an assessment was made the component was reloaded from the server. In this table, the number of assessments has therefore been subtracted from the number of requests for components. It can be seen that for the most part (70% of cases) searchers viewed 1 component and assessed it (or

Table 8. Relevance assessments distributed on access modes for the B tasks

Relevance	nextprev	rankedlist	structure	total
A	1	38	26	65
B	-	16	12	28
C	-	4	4	8
D	-	11	8	19
E	-	21	15	36
F	-	21	7	28
G	-	10	2	12
H	2	23	8	33
I	1	45	33	79
J	1	142	48	191
U	-	4	4	
Total	5	335	163	503

viewed two and didn't assess any), and then moved on to a new document rather than continuing the navigation within the same document.

A more in-depth analysis of the data will be performed with the aim to further break down user browsing behaviour within an accessed document. From informal comments made by searchers, and from an initial observation of the log data, one possible reason for the low degree of interaction with documents and their components was overlap. Searchers generally recognised overlapping components, and found them an undesirable "feature" of the system. Through more detailed analysis of the logs we can determine how searchers behaved when the system returned overlapping components.

3.3 Detailed Browsing Behaviour

A detailed analysis on the browsing behaviour of searchers is currently underway. The main aim of this analysis is to determine how users browsed within each document they

Table 9. Relevance assessments distributed on access modes for the C tasks

Relevance	nextprev	rankedlist	structure	total
A	-	43	18	61
B	-	25	11	36
C	-	11	2	13
D	-	30	15	45
E	-	34	27	61
F	-	26	12	38
G	-	13	7	20
H	-	35	12	47
I	-	60	20	80
J	-	64	20	84
U	-	4		4
Total	0	345	144	489

Table 10. Overall browsing behaviour within the same document: number of components viewed

	B	C	Total	B	C
1	406	394	800	69.0%	71.6%
2	93	84	177	15.8%	15.3%
3	47	39	86	8.0%	7.1%
4	23	9	32	3.9%	1.6%
5	13	8	21	2.2%	1.5%
6	2	4	6	0.3%	0.7%
7	2	5	7	0.3%	0.9%
8	1	1	2	0.2%	0.2%
9	1		1	0.2%	0.0%
10		1	1	0.0%	0.2%
11		2	2	0.0%	0.4%
12		1	1	0.0%	0.2%
13		1	1	0.0%	0.2%
14		1	1	0.0%	0.2%
Total	588	550	1138	100%	100%

visited, and how their browsing actions correlated with their relevance assessments. More specifically, we aim to look into the relationship of the relevance assessments' dimensions to whether searchers browse to more specific or more general components in the document tree, whether they browse to components of the same depth or whether they return to the ranked list of components. For example, we could see where users would browse to after they have assessed a component as "Very useful and fairly specific", and also how they would assess further documents along the browsing path.

This detailed analysis, together with the analysis on the overlapping components, can yield results that can be useful for the development of metrics that may take into account actual indications of user behaviour.

4 Conclusions

In this paper we have described the motivation and aims, and the methodology of the INEX 2004 interactive track. We have also presented some initial results gathered from user questionnaires and system logs.

We are currently performing a more detailed analysis of the gathered data, with the aim to establish patterns of browsing behaviours and to correlate them to the assessments of the visited document components. This analysis can also provide insight as to whether there are different browsing behaviours for the two different task categories included in the study. We expect that the results of this analysis will lead to the development of effectiveness metrics based on observed user behaviour.

References

1. Finesilver, K., Reid, J.: User behaviour in the context of structured documents. In: Sebastiani, Fabrizio (ed.), Advances in information retrieval: Proceedings of the 25th European conference on IR research, ECIR 2003. (2003) 104–199

2. Kazai, G.: Report of the INEX 2003 metrics working group. In Fuhr, N., Lalmas, M., Malik, S., eds.: INitiative for the Evaluation of XML Retrieval (INEX). Proceedings of the Second INEX Workshop. Dagstuhl, Germany, December 15–17, 2003. (2004) 184–190
3. Tombros, A., Ruthven, I., Jose, J.: How users assess web pages for information-seeking. Journal of the American Society for Information Science and Technology **56** (2005) 327–344
4. Borlund, P.: Evaluation of interactive information retrieval systems. PhD thesis, Royal School of Library and Information Sciences, Copenhagen, Denmark (2000)
5. Borlund, P.: The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Information Research:an international electronic journal **8** (2003) 1–38

A Task Descriptions

A.1 Task Category: Background (B)

Task ID: B1
You are writing a large article discussing virtual reality (VR) applications and you need to discuss their negative side effects. What you want to know is the symptoms associated with cybersickness, the amount of users who get them, and the VR situations where they occur. You are not interested in the use of VR in therapeutic treatments unless they discuss VR side effects.

Task ID: B2
You have tried to buy & download electronic books (ebooks) just to discover that problems arise when you use the ebooks on different PC's, or when you want to copy the ebooks to Personal Digital Assistants. The worst disturbance factor is that the content is not accessible after a few tries, because an invisible counter reaches a maximum number of attempts. As ebooks exist in various formats and with different copy protection schemes, you would like to find articles, or parts of articles, which discuss various proprietary and covert methods of protection. You would also be interested in articles, or parts of articles, with a special focus on various disturbance factors surrounding ebook copyrights.

A.2 Task Category: Background (C)

Task ID: C1
You have been asked to make your Fortran compiler compatible with Fortran 90, and so you are interested in the features Fortran 90 added to the Fortran standard before it. You would like to know about compilers, especially compilers whose source code might be available. Discussion of people's experience with these features when they were new to them is also of interest.

Task ID: C2
You are working on a project to develop a next generation version of a software system. You are trying to decide on the benefits and problems of implementation in a number of programming languages, but particularly Java and Python. You would like a good comparison of these for application development. You would like to see comparisons

of Python and Java for developing large applications. You want to see articles, or parts of articles, that discuss the positive and negative aspects of the languages. Things that discuss either language with respect to application development may be also partially useful to you. Ideally, you would be looking for items that are discussing both efficiency of development and efficiency of execution time for applications.