Overview of the INEX 2008 Interactive Track

Nils Pharo[1], Ragnar Nordlie[1] and Khairun Nisa Fachry[2]

[1]Faculty of Journalism, Library and Information Science, Oslo University College, Norway
nils.pharo@jbi.hio.no, ragnar.nordlie@jbi.hio.no
[2] Archives and Information Studies, University of Amsterdam, the Netherlands
k.n.fachry@uva.nl

Published as: Pharo, N., Nordlie, R. & Fachry, K.N. (2009). Overview of the INEX 2008 Interactive Track. Lecture Notes in Computer Science, 5631, 300-313.

Abstract. This paper presents the organization of the INEX 2008 *interactive track*. In this year's iTrack we aimed at exploring the value of element retrieval for two different task types, fact-finding and research tasks. Two research groups collected data from 29 test persons, each performing two tasks. We describe the methods used for data collection and the tasks performed by the participants. A general result indicates that test persons were more satisfied when completing research task compared to fact-finding task. In our experiment, test persons regarded the research task easier, were more satisfied with the search results and found more relevant information for the research tasks.

1 Introduction

The INEX interactive track (iTrack) is a cooperative research effort run as part of the INEX Initiative for the Evaluation of XML retrieval [1]. The overall goal of INEX is to experiment with the potential of using XML to retrieve relevant parts of documents through the provision of a test collection of XML-marked Wikipedia articles. The main body of work within the INEX community has been the development and testing of retrieval algorithms. Interactive information retrieval (IIR) [2] aims at investigating the relationship between end users of information retrieval systems and the systems. This aim is approached partly through the development and testing of interactive features in the IR systems and partly through research on user behavior in IR systems. In the INEX iTrack the focus has been on how end users react to and exploit the potential of IR systems that facilitate the access to *parts* of documents in addition to the full documents.

The INEX interactive track (iTrack) was run for the first time in 2004 [3], repeated in 2005 [4] and again in 2006/2007 [5] (due to technical problems the tasks scheduled for 2006 were actually run in early 2007). Although there has been variations in task content and focus, some fundamental premises has been in force throughout:

- a common subject recruiting procedure
- a common set of user tasks and data collection instruments such as questionnaires
- a common logging procedure for user/system interaction
- an understanding that collected data should be made available to all participants for analysis

This has ensured that through a manageable effort, participant institutions have had access to a rich and comparable set of data on user background and user behavior, of sufficient size and level of detail to allow both qualitative and quantitative analysis. This has already been the source of a number of papers and conference presentations ([6], [7], [8], [9], [10], [11], [12]).

In 2008, we wanted to preserve as much of the "common effort" quality of the previous years as possible. We invited the participants to participate in a minimum experimental effort using the system and data provided and described below. Within the framework of the track, participants could then design their own investigations under certain constraints, such as:

- The collection of documents was the same as the one used for the INEX ad hoc retrieval task [13], i.e., in 2008 a collection of xml-coded Wikipedia articles.
- The IR system developed for the 2006 track was made available for the participants to use, either alone or in comparison with participants' own system(s).
- Each participating site was responsible for recruiting a minimum of 8 (but preferably more) test persons to participate in the study as searchers.
- The participants were required to make their data available to all participating groups, and describe their collection process and experimental procedure in a way which would make it possible for others to interpret and use the data.

2 Tasks

For the 2008 iTrack the experiment was designed with two categories of tasks, from each of which the searchers were instructed to select one out of three alternative search topics constructed by the track organizers. The original intention was to also give the searchers the opportunity to perform one self-generated task, but it was unfortunately not possible to implement this in our IR system. The two categories of tasks were, respectively, fact-findings tasks (category 1) and research tasks (category 2). The tasks were intended to represent information needs believed to be typical for Wikipedia users. In order to ensure a certain amount of user-system interaction, we also wanted the tasks to be so complex that searchers needed to access more than one individual article to solve them. In order to diminish system learning effect, the order of tasks performed by searchers was rotated by category.

The fact-finding tasks:

sto1. As a frequent traveler and visitor of many airports around the world you are keen on finding out which is the largest airport. You also want to know the criteria used for defining large airports.

sto2. The "Seven summits" are the highest mountains on each of the seven continents. Climbing all of them is regarded as a mountaineering challenge. You would like to know which of these summits were first climbed successfully.

sto3. In the recent Olympics there was a controversy over the age of some of the female gymnasts. You want to know the minimum age for Olympic competitors in gymnastics.

The research tasks:

sto4. You are writing a term paper about political processes in the United States and Europe, and want to focus on the differences in the presidential elections of France and the United States. Find material that describes the procedure of selecting the candidates for presidential elections in the two countries.

sto5. Every year there are several ranking lists over the best universities in the world. These lists are seldom similar. You are writing an article discussing and comparing the different ranking systems and need information about the different lists and what criteria and factors they use in their ranking.

sto6. You have followed the news coverage of the conflict between Russia and Georgia over South Ossetia. You are interested in the historic background for the conflict and would like to find as much information about it as possible. In particular you are interested in material comparing this conflict with the parallel border conflict between Georgia and Abkhazia.

3 Participating groups

Originally 7 groups expressed their interest in participating in the i-Track experiments. Unfortunately, in the end only two groups were able to perform experiments; University of Amsterdam and Oslo University College. Fifty-six sessions, 14 in Amsterdam and 42 in Oslo, performed by 29 test persons were recorded successfully (i.e. without system failure and with completed questionnaires).

4 Research design

4.1 Search system

The experiments were conducted on a java-based retrieval system built within the Daffodil framework [14], which resides on a server at and is maintained by the University of Duisburg. The search system interface is quite similar to the one used in the 2005 and 2006 i-Tracks.

The system returns elements of varying granularity (full Wikipedia articles, sections or sub-sections of articles) based on the hierarchical xml-coded document structure. Figure 1 shows the result list interface of the program. In the top left corner is the query box, below it we see the result list. Relevant elements are grouped by document in the result list and up to three high ranking elements are shown per document. To help searchers select query terms, the system has a related term feature which presents the searcher with a set of potential query terms, generated through analysis of term frequency in the top-ranked elements. These appear in a box showing terms related to the current query. Using mouse-over, searchers can view the context from which the related terms were generated.

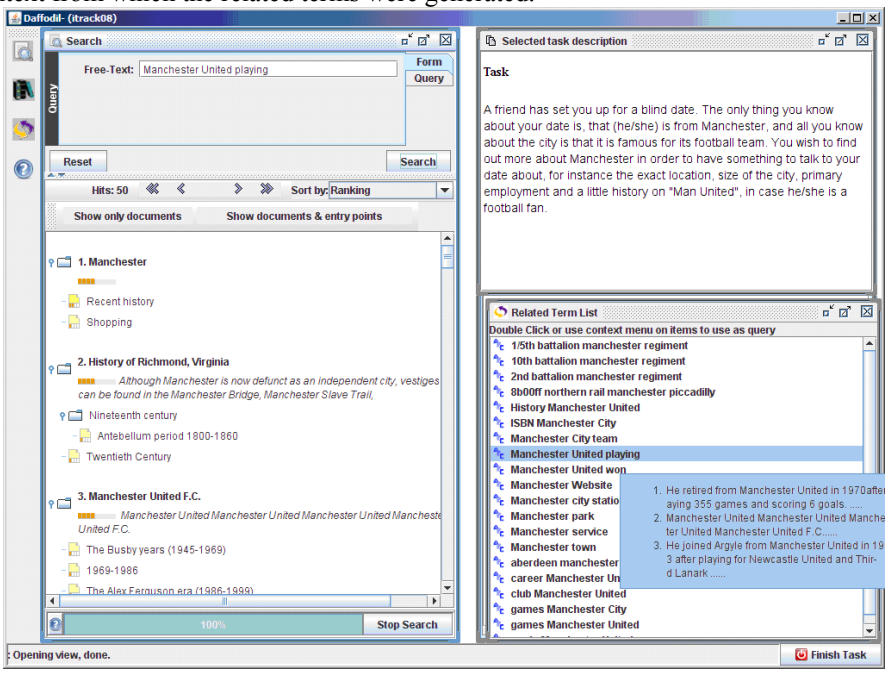

Fig. 1. Daffodil result list view

When a searcher clicks on the result list to examine a document, the system enters document view mode, where the entire full text of the document is shown, with background highlighting for high ranking elements (Figure 2). In addition to this, the document view screen shows a Table of Contents generated from the XML formatting

of the documents. From the ToC, the searcher can choose individual sections and subsections for closer examination. In the ToC, the system's relevance estimation is also indicated through color-coding of relevant elements. In addition, the ToC shows elements that the searcher has viewed (indicated by an eye -) and/or relevance assessed (coded as shown in Figure 3).

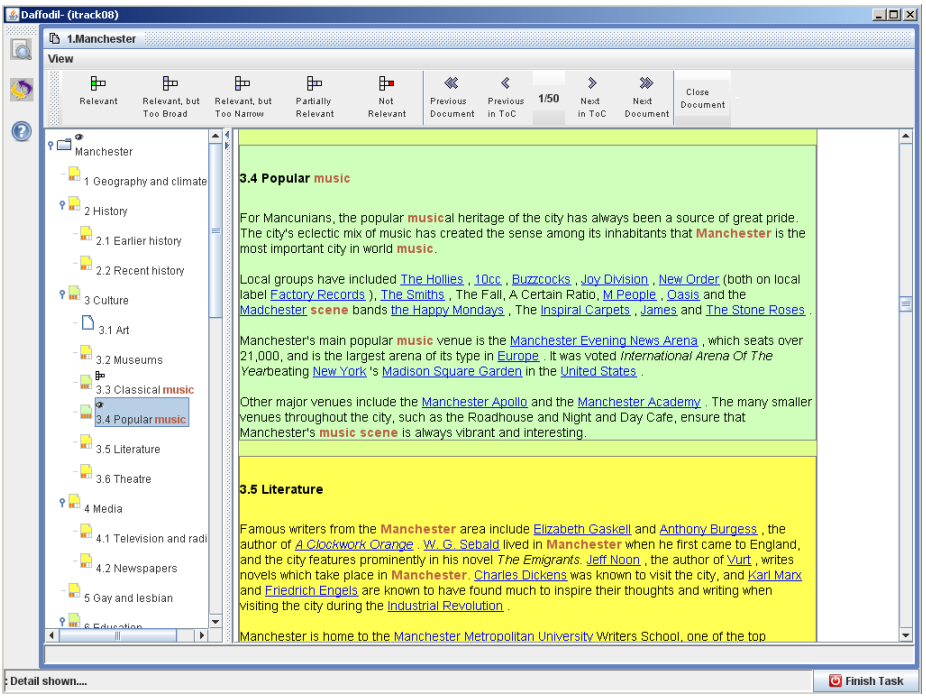

Fig. 2. Document view

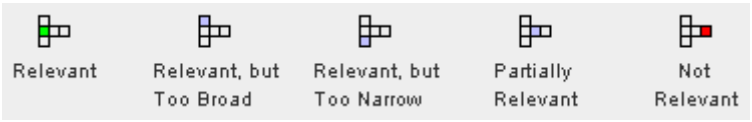

Fig. 3. Relevance scores

4.2 Document corpus

The document corpus used was the same as the one used in the 2006 i-track and in the other 2008 INEX tracks. It consists of more than 650,000 encyclopedia articles extracted from Wikipedia [15]. The articles are structurally formatted in XML.

4.3 Online questionnaires

During the course of the experiment, searchers were issued brief online questionnaires to support the analysis of the log data. Before the search tasks were introduced, the searchers were given a pre-experiment questionnaire, with demographic questions such as searchers' age, education and experience in information searching, particularly in searching and using Wikipedia. Each search task was preceded with a pre-task questionnaire, which concerned searchers' perceptions of the difficulty of the search task, their familiarity with the topic etc. After each task, the searcher was asked to fill out a post-task questionnaire. The intention of the post-task questionnaire was to learn about the searchers' use of and their opinion on various features of the search system, in relation to the task they had just completed. The experiment was closed with a post-experiment questionnaire, which elicited the searchers' general opinion of the search system. The responses to the questionnaires were logged in a database.

4.4 Relevance assessments

The system was designed to have searchers assess the relevance of each item they looked at. These could be either full articles or article elements. The relevance scale (see fig. 3) was similar to the one used in the 2006 interactive track, based on work by Pehcevski [16]. It aims to balance the need for information on the perceived granularity of retrieved elements and their degree of relevance, and is intended to be simple and easy to visualize [5]. The system did not oblige searchers to perform relevance judgments, but in the instructions for the experiment they were told to "select an assessment for each viewed piece of information with regards to how you consider it to be of help in solving the task." Searchers were not given any more specific instructions on how to perform the relevance judgments; they were, for instance, not required to view each retrieved element as independent from other components viewed. Experiences from user studies (e.g. [17]) clearly show that users learn from what they see during a search session. To impose a requirement on searchers to discard this knowledge were thought to create an artificial situation and restrain the searchers from interacting with the retrieved elements in a natural way.

Five different relevance scores were defined. The scores express two aspects or dimensions in relation to solving the task:

1. How much **relevant information** does the part of the document contain? It may be *highly relevant, partially relevant* or *not relevant*.
2. How much **context is needed** to understand the element? It may be *just right, more* or *less*.

This is combined into the five scores:

Relevant, but too broad, contains relevant information, but also a substantial amount of other information.

Relevant, contains highly relevant information, and is just the right in size to be understandable.

Relevant, but too narrow, contains relevant information, but needs more context to be understood.

Partially relevant, has enough context to be understandable, but contains only partially relevant information.

Not relevant, does not contain any relevant information that is useful for solving the task.

4.5 Logging

All search sessions were logged and saved to a database. The logs registered and time stamped the events in the session and the actions performed by the searcher, as well as the responses from the system.

5 Experimental Procedure

Each experiment was performed following the standard procedure outlined below. Steps 7 to 10 were repeated for each of the two tasks performed by the searcher. The tasks were automatically assigned according to a Latin square design to secure a balanced distribution of the order of the research and fact-finding tasks.

1. Experimenter briefed the searcher, and explained format of study. The searcher read and signed the Consent Form.
2. The experimenter logged the searchers into the experimental system. Tutorial of the system was given with a training task provided by the system. The experimenter handed out and explained the system features document.
3. Any questions answered by the experimenter.
4. The control system administered the pre-experiment questionnaire.
5. Topic descriptions for the first task category administered, and a topic selected
6. Pre-task questionnaire administered.
7. Task began by clicking the link to the search system. Maximum duration for a search was 15 minutes, at which point the system issued a “timeout” warning. Task ended by clicking the “Finish task” button.
8. Post-task questionnaire administered.
9. Steps 5-8 repeated for the second task.
10. Post-experiment questionnaire administered.

6 Data analysis

In this section, we summarize our preliminary analysis of the questionnaire data and the transaction log files. More detailed analyses will be the subject of further research from the participating institutions.

Table 1. Distribution of tasks and sessions

Task Type	Task	Sessions
Fact-finding	Sto1	13
	Sto2	8
	Sto3	5
Research	Sto4	9
	Sto5	9
	Sto6	12
Total		56

Table 1 shows the distribution of tasks and sessions, due to a technical error one searcher performed two research tasks and one searcher performed only one task (also a research task) thus it is not a completely even distribution of task types (26 fact-finding tasks and 30 research tasks).

6.1 Questionnaire data

Questionnaire results reported in this report are based on the data of test persons who completed the questionnaire.

Pre-Experiment Questionnaire

A total number of 27 test persons completed the questionnaire (9=Male, 18=Female). Test persons had a mean age of 30.33 years and with the exception of six test persons, all were students. Test persons' mean experience with searching for information using the Web was 8.22 years. When asked about how often they search, our test persons' mean search experience using digital libraries was 3.60, using search engines was 4.81, and using Wikipedia was 3.81 (where 1=never, 2 = once or twice a year, 3 = once or twice a month, 4 = once or twice a week and 5 = once or more times a day).

As we were using Wikipedia, we administered test persons' experiences with Wikipedia in detail. First, we asked about the test persons' search purposes with Wikipedia. Out of 27 test persons, 25 of them mentioned that they used Wikipedia for fact-finding purposes, none of them used Wikipedia for decision making, 10 test persons used Wikipedia for research and 9 test persons used Wikipedia for entertainment. When asked if they generally found what they were looking for when using Wikipedia, they responded positively (their mean experience was 3.96), and when asked if they trust the information in Wikipedia, subjects mean experience was

3.41 (where 1=strongly disagree, 2=disagree, 3=not sure, 4=agree and 5=strongly agree). Lastly, our pre-experiment questionnaire result indicated that only 1 out of 27 test persons mentioned that he or she occasionally has edited articles in Wikipedia and none of our users ever have created new articles in Wikipedia.

Pre-Task Questionnaire

Table 2. Pre-task questionnaire, with answers on a 5-point scale (1-5)

Q2.1:	How familiar are you with the topic of the search task?
Q2.2:	How interesting do you find the topic of the search task?
Q2.3:	How easy do you think it will be to find information for this task

Table 3. Pre-task responses on searching experience: mean scores and standard deviations (in brackets)

Type	Q2.1	Q2.2	Q2.3
All tasks	1.96 (0.78)	3.43 (0.73)	3.22 (0.68)
Fact Finding	1.81 (0.84)	3.26 (0.68)	3.59 (0.53)
Research	2.11 (0.74)	3.59 (0.69)	2.85 (0.72)

Each task was preceded with a pre-task questionnaire, collecting information regarding test persons' familiarity, level of interest and easiness of the search topic. Table 2 shows the items asked in the pre-task questionnaire. The answer categories used a 5-point scale (1=not at all, 3=somewhat and 5=extremely). Test persons' responses are presented in table 3.

As shown in table 3, the research task was rated slightly higher compared to fact-finding task in terms of test person's familiarity with the topic (Q2.1) and level of interest (Q2.2) of the search task. Only in terms of perceived easiness to find information for the task (Q2.3), the fact-finding task was rated higher.

Post-Task Questionnaire

Table 4 shows the items asked in the post-task questionnaire. The answer categories used a 5-point scale (1=not at all, 3=somewhat and 5=extremely). Test persons' responses are summarized in Table 5. If we look at the responses over all tasks, the average response varies from 2.83 to 4.46 signaling that the test persons rated the tasks positively.

We also looked at the responses for each task type. As shown in table 5, for all questions asked with the exception of Q3.1 and Q3.8, the research task was rated higher than the fact-finding task. Here, we see that test persons understood both tasks very well (Q3.1). Fact-finding received higher responses on average, which makes sense given the nature of the simulated tasks and thereby confirms that the chosen simulated tasks represent the particular task types. The research task was regarded easier (Q3.2) and more similar to the searching task that our test persons typically perform (Q3.3), compare to the fact-finding task. This may be a result of our selection of test persons who all had an academic education. Moreover, test persons were more satisfied with the search results provided by the system (Q3.6) for the research task. A

possible explanation is that the research tasks are more open-ended than the fact-finding tasks where test persons need to find specific and precise answers. Hence, additional material provided by the system may be more useful in the research task context. This explanation is supported by the response when asked about the relevancy of the found information (Q3.7). Test persons believed that they found more relevant results for the research tasks. This finding is also coherent with the relevance assessment results where searchers found more articles and more elements to be relevant when completing research tasks compare to when they performed fact-finding tasks (see Section 6.2).

Table 4. Post-task questionnaire, with answers on a 5-point scale (1-5).

Q3.1:	How understandable was the task?
Q3.2:	How easy was the task?
Q3.3:	To what extent did you find the task similar to other searching tasks that you typically perform?
Q3.4	Was it easy to perform the search for this task?
Q3.6:	Are you satisfied with your search results?
Q3.7:	How relevant was the information you found?
Q3.8:	Did you have enough time to do an effective search?
Q3.9:	How certain are you that you completed the task?
Q3.10	How well did the system support you in this task?*

Table 5. Post-task responses on searching experience: mean scores and standard deviations (in brackets)

Type	Q3.1	Q3.2	Q3.3	Q3.4	Q3.6	Q3.7	Q3.8	Q3.9	Q3.10
All tasks	4.46 (0.64)	3.13 (1.27)	3.46 (1.04)	3.31 (1.06)	3.02 (1.51)	3.50 (1.28)	3.04 (1.45)	2.83 (1.46)	3.02 (1.22)
Fact Finding	4.63 (0.56)	3.00 (1.47)	3.30 (1.10)	3.19 (1.11)	2.56 (1.63)	3.07 (1.38)	3.07 (1.57)	2.63 (1.64)	2.70 (1.05)
Research	4.30 (0.67)	3.26 (1.06)	3.63 (0.97)	3.44 (1.01)	3.48 (1.25)	3.93 (1.04)	3.00 (1.36)	3.04 (1.26)	3.33 (0.91)

Next, we look at the time test persons spent on each task. On the question of whether there was enough time for an effective search (Q3.8), responses for the fact-finding tasks were higher than for the research tasks. This is also consistent with the log result where test persons spent less time completing fact-finding tasks compared to research tasks (see Section 6.2). This means that test persons had enough time for the fact-finding task, but they stopped searching before the maximum allocated time ran out. This could be because the system did not support them well enough in finding relevant results (Q3.10) or they expected the system to do better in retrieving relevant results (Q3.7) for fact-finding tasks. This is consistent with the assessment of task completion (Q3.9) where, on average, test persons were less certain that they completed the fact-finding task compared to the research task. Also note that the standard deviations for fact-finding tasks for almost all questions are larger than for the research tasks. A possible explanation is again that several test persons were not satisfied with the results they found when completing the fact-finding task.

6.2 Log statistics

In total 118 assessments were made of full articles, Table 6 shows the distribution of assessment on the different relevance levels.

Table 6. Article relevance assessments

Fully relevant	Relevant, but too broad	Relevant, but too narrow	Partially relevant	Not relevant
45 (38 %)	14 (12 %)	12 (10 %)	17 (14 %)	30 (25 %)

In Table 7, we see relevance distribution of articles for each topic, the results show that the sessions generated by task sto6 (on the South Ossetia conflict), which is the most popular research task, has returned more than half of the articles found to be fully relevant. Even more interesting to see is that sessions dealing with the most popular fact-finding task (sto1 – large airports) has not returned any fully relevant articles.

Table 7. Distribution of article relevance assessments per task

Topic	Fully relevant	Relevant, but too broad	Relevant, but too narrow	Partially relevant	Not relevant
sto1	0	2	6	5	11
	.0%	14.3%	14.3%	29.4%	34.4%
sto2	2	0	0	2	2
	4.4%	.0%	.0%	11.8%	6.3%
sto3	2	0	0	4	11
	4.4%	.0%	.0%	23.5%	34.4%
sto4	7	1	2	1	3
	15.6%	7.1%	15.4%	15.4%	9.4%
sto5	9	1	3	4	3
	20.0%	7.1%	23.1%	23.5%	9.4%
sto6	25	10	2	1	2
	55.6%	71.4%	15.4%	5.9%	6.3%
Total	45	14	13	17	32
	100.0%	100.0%	100.0%	100.0%	100.0%

Table 8 shows the distribution of relevance assessments on element level, i.e. assessments of sections and subsections. Interestingly we also see that task sto6 also on the element level has returned the highest number of fully relevant scores and that sto1 only has returned 3 fully relevant elements.

Table 8. Distribution of element relevance assessments per task

Topic	Fully relevant	Relevant, but too broad	Relevant, but too	Partially relevant	Not relevant

			narrow		
sto1	3	2	1	3	5
	2.7%	25.0%	3.3%	7.5%	10.6%
sto2	6	1	7	1	1
	5.3%	12.5%	23.3%	2.5%	2.1%
sto3	0	0	0	0	1
	.0%	.0%	.0%	.0%	2.1%
sto4	5	2	4	7	3
	4.4%	25.0%	13.3%	17.5%	6.4%
sto5	44	1	5	18	33
	38.9%	12.5%	16.7%	45.0%	70.2%
sto6	55	2	13	11	4
	48.7%	25.0%	43.3%	43.3%	8.5%
Total	113	8	30	40	47
	100.0%	100.0%	100.0%	100.0%	100.0%

We have performed further analysis to investigate if there are any significant differences between the two task types. A T-test shows a significant difference between fully relevant assessment on both article (p=0.000) and element level (P=0.011) when comparing fact-findings tasks and research tasks, but then one needs to be aware of the heavy influence of relevance assessments for tasks sto1 and sto6. For fact-finding tasks searchers found 0.15 fully relevant articles per session and 0.35 fully relevant elements, compared to 1.37 fully relevant articles and 3.47 elements per session for research tasks. Also fact-finding sessions resulted in significantly more non relevant articles (1.197 compared to 0.583 for research tasks). This supports the findings from the questionnaire analysis that searchers were more familiar with the research tasks and found them easier to solve, and also that they believed they found more relevant information for the research tasks.

Table 9. Queries per task

	Task type	N	Mean
Number of queries	Fact	26	5.88
	Research	30	4.83

Table 10. Time per task

	Task type	N	Mean
Time in seconds	Fact	26	653.15
	Research	30	767.10

We have also compared the task types with respect to number of queries (Table 9) performed and time invested (Table 10). As can be seen the searchers performed more queries in fact-finding sessions but, but spent more time to solve research tasks. In other words research task sessions are characterized by searchers being more thorough in their interaction with the individual article/element. A T-test did not

report significant difference between the two task categories in these matters, but the mean time per task was very close to being significant (p=0.064).

7 Conclusions

We have reported the experimental design of the 2008 Inex interactive track and the analysis of data related to the difference between searchers performing fact-finding and research tasks. Although the number of participating institutions was low, we have been able to collect a set of data that shows interesting results related to the two task categories.

In general, searchers were more satisfied when completing the research task compared to fact-finding task. We found that test persons regarded the research task easier, were more satisfied with the search result and found more relevant information for the research task. This is plausibly related to the task type, where test persons regard more information as relevant or useful when searching for a more open-ended research task. Fact-finding tasks require a more specific and precise answer, which may diminish the additional value of exploring a wide range of search results.

This finding is consistent with the relevance assessment results where searchers found more relevant articles and elements when completing the research task compared to the fact-finding task. Also fact-finding sessions resulted in significantly more non-relevant articles than research sessions. Test persons reported that they were less certain that they had completed the fact-finding task compared to the research task.

A general result seems to be that the system was better at supporting research tasks than fact-finding tasks. This is particularly interesting since the participants claimed to use Wikipedia more for fact-finding than for research tasks.

Acknowledgments.

We would like to thank Ingo Frommholz, Norbert Fuhr, Claus-Peter Klas and Saadia Malik from the University of Duisburg-Essen for their administration of the Daffodil system. Khairun Nisa Fachry was supported by the Netherlands Organization for Scientific Research (NWO) under grant # 639.072.601.

References

[1] Malik, S., Trotman, A., Lalmas, M. & Fuhr, N. (2007): Overview of INEX 2006. In: Fuhr, N., Lalmas, M. and Trotman, A. eds. *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 2006*. Berlin: Springer, p. 1-11.

[2] Ruthven, I. (2008): Interactive Information Retrieval. In: Annual Review of Information Science and Technology, 42, p. 43-91.

[3] Tombros, A., Larsen, B. and Malik, S. (2005): The Interactive Track at INEX 2004. In: Fuhr, N., Lalmas, M., Malik, S. and Szlávik, Z. eds. *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004*. Berlin: Springer, p. 410-423

[4] Larsen, B., Malik, S. and Tombros, A. (2006): The interactive track at INEX 2005. In: Fuhr, N., Lalmas, M., Malik, S. and Kazai, G. eds. *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005*. Berlin: Springer, p. 398-410.

[5] Larsen, B., Malik, S. & Tombros, A. (2007): The Interactive track at INEX 2006. In: Fuhr, N., Lalmas, M. and Trotman, A. eds. *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 2006*. Berlin: Springer, p. 387-399.

[6] Pharo, N. & Nordlie, R. (2005): Context Matters: An Analysis of Assessments of XML Documents. In: F. Crestani and I. Ruthven eds. *Information Context: Nature, Impact, and Role: 5th International Conference on Conceptions of Library and Information Sciences, CoLIS 2005, Glasgow, UK, June 4-8, 2005*. Berlin: Springer, p. 238-248.

[7] Hammer-Aebi, B., Christensen, K. W., Lund, H. and Larsen, B. (2006): Users, structured documents and overlap: interactive searching of elements and the influence of context on search behaviour. In: Ruthven, I. et al. eds. *Information Interaction in Context : International Symposium on Information Interaction in Context : IIIiX 2006 : Copenhagen, Denmark, 18-20 October, 2006 : Proceedings.* Copenhagen: Royal School of Library and Information Science, p. 80-94.

[8] Malik, S., Klas, C.-P., Fuhr, N., Larsen, B. and Tombros, A. (2006): Designing a user interface for interactive retrieval of structured documents: lessons learned from the INEX interactive track? In: Gonzalo, J. et al. eds. *Research and Advanced Technology for Digital Libraries, 10th European Conference, ECDL 2006*. Alicante, Spain, September 17-22, 2006, Proceedings. Berlin: Springer,

[9] Kim, H. & Son, H. (2006): Users Interaction with the Hierarchically Structured Presentation in XML Document Retrieval. In: Fuhr, N., Lalmas, M., Malik, S. & Kazai, G. eds. *Advances in XML Information Retrieval and Evaluation: 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005.* Berlin: Springer, p. 422-431.

[10] Kazai, G. & Trotman, A (2007): Users' perspectives on the Usefulness of Structure for XML Information Retrieval. In: Dominich, S. & Kiss, F. eds. *Proceedings of the 1st International Conference on the Theory of Information Retrieval*. Budapest: Foundation for Information Society, p. 247-260.

[11] Larsen, B., Malik, S & Tombros, A. (2008): A Comparison of Interactive and Ad-Hoc Relevance Assessments. In: Fuhr, N., Kamps, J., Lalmas, M. & Trotman, A. eds. *Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany, December 17-19, 2007.* Berlin: Springer, p. 348-358.

[12] Pharo, N. (2008): The effect of granularity and order in XML element retrieval. Information Processing and Management. 44(5), 1732-1740.

[13] Kamps, J., Geva, S., Trotman, A., Woodley, A. & Koolen M. (2009). Overview of the INEX 2008 Ad Hoc Track. In: *Proceedings from the 7th*

International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009. Berlin: Springer

[14] Fuhr, N., Klas, C.P., Schaefer, A. & Mutschke, P. (2002): Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, p. 597-612.

[15] Denoyer, L. & Gallinari, P. (2006): The Wikipedia XML corpus. In: SIGIR Forum. 40(1), p. 64-69

[16] Pehcevski, J. (2006): Relevance in XML retrieval: the user perspective. In: Trotman, A. and Geva, S. eds. *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology : Held in Seattle, Washington, USA, 10 August 2006*. Dunedin (New Zealand): Department of Computer Science, University of Otago, p. 35-42.

[17] Pharo, N. (2002): The SST Method Schema: a tool for analyzing work task-based Web information search processes. Doctoral Thesis. University of Tampere