

The Interactive Track at INEX 2005

Birger Larsen¹, Saadia Malik², and Anastasios Tombros³

¹ Dept. of Information Studies, Royal School of LIS, Copenhagen, Denmark
blar@db.dk

² Fak. 5/IIS, Information Systems, University of Duisburg-Essen,
Duisburg, Germany
malik@is.informatik.uni-duisburg.de

³ Dept. of Computer Science, Queen Mary University of London, London, UK
tassos@dcs.qmul.ac.uk

Abstract. In its second year, the Interactive Track at INEX focused on addressing some fundamental issues of interactive XML retrieval: is element retrieval useful for searchers, what granularity of elements do searchers find more useful, what applications for element retrieval can be viable in interactive environments, etc.. In addition, the track also expanded by offering an alternative document collection, by including two additional tasks, and by attracting more participating groups: A total of 11 research groups and 119 test persons participated in the three different tasks that were included in the track. In this paper, we describe the main issues that the Interactive Track at INEX 2005 attempts to address and the methodology and tasks that were used in the track.

1 Introduction

The overall motivation for the Interactive Track at INEX is 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.

One of the major outcomes of the Interactive Track in 2004 was the need to investigate methods that can be supportive during the search process based on features extracted from the XML formatting [1, 2]. Problems that might be solved using such methods include overlapping components, and the presentation of retrieved elements in the hit list.

In the system that was offered by the track in 2005 these two issues were addressed. This offered us the opportunity to study how overall user search behaviour was affected by these changes when compared to the behaviour observed in 2004.

In addition, the following aims were addressed in 2005 following the recommendations of the INEX Methodology Workshop at the Glasgow IR Festival¹:

- To elicit user perceptions of what is needed from an XML retrieval system. The aim is to see whether element retrieval is what users need: Does element retrieval make sense at all to users, do users prefer longer components, shorter components or whole documents, would they rather have passages than elements, etc.

¹ See <http://www.cs.otago.ac.nz/inexmw/> for the proceedings and presentation slides.

- To identify an application for element retrieval. This year, a mixture of topics that were simulated work tasks [3] (based on topics from the ad hoc track) and information needs formulated by the test persons themselves. The aim of including the latter was to enable studies of what characterises the tasks users formulate, and to see what kinds of applications users might need an element retrieval system for. A total of 121 such topics derived from the test persons were collected for further analysis.
- To introduce an alternative document collection with the Lonely Planet collection as an optional task in order to broaden the scope of INEX and to allow test persons with different backgrounds (e.g. educational) to participate.

The format of the Interactive Track in 2005 was deliberately of an exploratory nature, and has relatively broad aims rather than addressing very specific research questions. Element retrieval is still in its infancy and many basic questions remain unanswered as shown by the discussions at the IR Festival. Aside from the automatic and detailed logging of test persons as used last year, more emphasis was placed on producing qualitative results. Many of the aims stated above were therefore dealt with through careful interviewing and detailed questionnaires. A total of three tasks were available to the track participants: one compulsory task that all participants had to fulfil with a minimum number of test persons, and two optional tasks. These tasks combined several element retrieval systems, topic types and XML collections. By providing a multitude of different perspectives it is our hope that the Interactive Track can aid in illuminating some of the core issues in element retrieval.

The remainder of the paper is organised as follows: The three tasks are described briefly in Section 2, followed by details of the participating groups in Section 3. In depth descriptions of Task A and Task C are given in Sections 4 and 5 respectively, whereas Task B is only described briefly in Section 2. Concluding remarks are given in Section 6.

2 Tasks in the INEX 2005 Interactive Track

2.1 Task A - Common Baseline System with IEEE Collection

In this task each test person searched three topics in the IEEE collection: Two simulated work tasks provided by the organisers, and one formulated by the test person herself in relation to an information need of her own. The baseline system used by all participants was a java-based element retrieval system built within the Daffodil framework², and was provided by the track organisers. It has a number of improvements over last year's baseline system, including handling of overlaps, better element summaries in the hit list, a simpler relevance scale, and various supportive interface functionalities. Task A was compulsory for all participating groups with a minimum of 6 test persons.

2.2 Task B - Participation with Own Element Retrieval System

This task allowed groups who have a working element retrieval system to test their system against a baseline system. Groups participating in Task B were free to choose

² See <http://www.is.informatik.uni-duisburg.de/projects/daffodil/index.html.en>

between the IEEE collection or the Lonely Planet collection, and had a large degree of freedom in setting up the experiment to fit the issues they wanted to investigate in relation to their own system. If the IEEE collection was used Daffodil was offered as baseline system. For the Lonely Planet collection a baseline system was kindly provided by the Contentlab at Utrecht University³. The recommended experimental setup was very close to that of Task A, with the main difference that simulated work tasks should be assigned to test persons rather than freely chosen. This in order to allow for direct comparisons between the baseline system and the local system.

Task B was optional for those groups who had access to their own element retrieval system, and was separate from task A. Thus additional test persons needed to be engaged for task B. See [7] for an example of an experimental setup used in Task B.

2.3 Task C - Searching the Lonely Planet Collection

This task allowed interested groups to carry out experiments with the Lonely Planet collection. Each test person searched four topics which were simulated work tasks provided by the organisers. The system (B³-SDR) provided by Utrecht University was used in this task. The system is a fully functional element retrieval system that supports several query modes. Task C was optional for those groups who wished to do experiments with the new collection, and was separate from task A and B. Thus additional test persons needed to be engaged for task C. Note that the Lonely Planet collection allows for test persons that do not have a computer science background (in contrast to the IEEE CS collection used in Task A).

Detailed experimental procedures including questionnaires and interview guides for all three tasks were provided to the participants. In addition, a specification of a minimum logging format was provided for local systems in Task B [8]. As for last year, minimum participation in the INEX Interactive Track did not require a large amount of work as the baseline system for Task A was provided by the track. The bulk of the time needed for participating groups was spent on running the experiments; approximately 2 hours per test person.

3 Participating Groups

A total of 12 research groups signed up for participation in the Interactive Track and 11 completed the minimum number of required test persons. Their affiliations and distribution on tasks are given in Table 1 below. All 11 groups participated in Task A with a total of 76 test persons searching on 228 tasks. Only one group, University of Amsterdam, participated in Task B with 14 test persons searching on 42 tasks. Four groups participated in Task C with 29 test persons searching 114 tasks. A total of 119 test persons from the 11 active participants took part in the Interactive Track. In comparison, in 2004, 10 groups took part with 88 test persons.

³ See <http://contentlab.cs.uu.nl/>

Table 1. Research groups participating in the Interactive Track at INEX2005

Research Group	Task A	Task B	Task C
	Test Persons (Topics)	Test Persons (Topics)	Test Persons (Topics)
CWI, University of Twente, The Netherlands	6 (18)	-	-
Kyungpook National University, Korea	12 (36)	-	-
Oslo University College, Norway	8 (24)	-	-
Queen Mary University of London, England	6 (18)	-	-
RMIT University, Australia	6 (18)	-	12 (48)
Royal School of LIS, Denmark	6 (18)	-	6 (24)
Rutgers University, USA	6 (18)	-	4 (16)
University of Amsterdam, The Netherlands	6 (18)	14 (42)	-
University of Duisburg-Essen, Germany	6 (18)	-	-
University of Tampere, Finland	8 (24)	-	-
Utrecht University, The Netherlands	6 (18)	-	7 (26)
Total	76 (228)	14 (42)	29 (114)

4 Task A

4.1 Document Corpus

The document corpus used in Task A was the 764 MB corpus of articles from the IEEE Computer Society's journals covering articles from 1995-2004 (version 1.8, merged new & old collection).

4.2 Relevance Assessments

The intention was that each viewed element should be assessed with regard to its relevance to the topic by the test person. This was, however, not enforced by the system as we believe that it may be regarded as intrusive by the test persons [4]. In addition, concerns have been raised that last year's composite two dimensional scale was far too complex for the test persons to comprehend [5, 6]. Therefore it was chosen to simplify the relevance scale, also in order to ease the cognitive load on the test persons. The scale used was a simple 3-point scale measuring the usefulness (or pertinence) of the element in relation to the test person's perception of the task:

- 2 – Relevant
- 1 – Partially Relevant
- 0 – Not Relevant

Please note that in contrast to the assessments made for the ad hoc track, there was no requirement on the test persons to view each retrieved element as independent from other viewed components. We have chosen not to enforce any rules in order to allow the test persons to behave as close as possible to what they would normally do.

For Task C we experimented with a slightly more complex relevance scale (see Section 6.2 below).

4.3 System

The baseline system used in Task A was a Java-based element retrieval system built within the Daffodil framework. The HyREX retrieval engine⁴ was used as backend in the baseline system.

Fig. 1 shows the query and results list interface of the baseline system. After entering a query and pressing “Search” a search progress indicator informed the test person about the number documents found. A related term list also appeared, suggesting alternative search terms (not shown). The results were presented as documents and in some cases, the system indicated which elements that might be most closely related to the query.

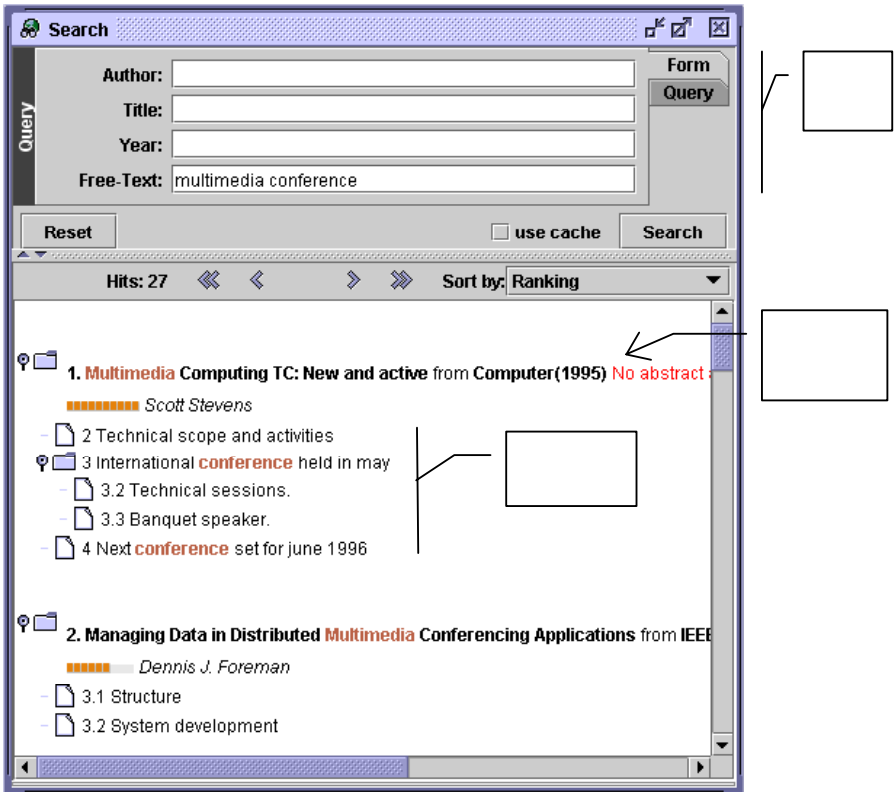
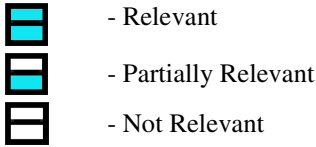


Fig. 1. Query box and result list display in the baseline system used in Task A

Double-clicking a document or an element opened this in a new window as shown in Fig. 2 below. This was split in two panes: one with a Table of Contents of the whole document, and one with the full text of the selected element. The selected

⁴ <http://www.is.informatik.uni-duisburg.de/projects/hyrex/>

element was displayed on the right. On the left, the Table of Contents indicated the currently viewed element, other retrieved elements, viewed and assessed elements. The relevance scale was implemented as simple icons to be clicked:



The logging in the baseline system was saved to a database for greater flexibility and stability. The log data comprises of one session for each topic the test person searches. The log for each session recorded the events in the session, both the actions performed by the test person and the responses from the system.

4.4 Tasks/Topics

In order to study the questions outlined in Section 1 above related to the needs for element retrieval systems and possible applications of such systems, both real and simulated information needs were used in Task A.

The test persons were asked to supply examples of own information needs. As it may be hard for the test persons to formulate topics that are covered by the collection, the test persons emailed two topics they would like to search for 48 hours before the experiment. The experimenters then did a preliminary search of the collection to determine which topic had the best coverage in the collection. The topics supplied by the test persons were not all well-suited to an element retrieval system, but they all had a valuable function as triggers for the structured interview where it was attempted to elicit user perceptions of what they need from an element retrieval system, and to identify possible applications for element retrieval. They may also be valuable for the formulation of topics for next year's track. Therefore, both topics were recorded and submitted as part of the results.

The simulated work tasks were derived from the CO+S and CAS INEX 2005 ad-hoc topics, ignoring any structural constraints. In order to make the topics comprehensible by other than the topic author, it was required that the ad hoc topics not only detail *what* is being sought for, but also *why* this is wanted, and in what *context* the information need has arisen. This information was exploited for creating simulated work task situations for Task A that, on the one hand will allow the test persons to engage in realistic searching behaviour, and on the other provide a certain level of experimental control by being common across test persons⁵.

For Task A, six topics were selected and modified into simulated work tasks. In last year's track we attempted to identify tasks of different types and to study the difference between them, but without great success. This year a simple partition has been made into two categories:

⁵ See the work of Borlund for more information on simulated work tasks, e.g. Borlund, 2003 (<http://informationr.net/ir/8-3/paper152.html>).

- General tasks (G category), and
- Challenging tasks (C category), which are more complex and may be less easy to complete.

In addition to their own information need, each test person chose one task from each category. This allows the topic to be more “relevant” and interesting to the test person. A maximum time limit of 20 minutes applied for each task. Sessions could finish before this if the test person felt they have completed the task.

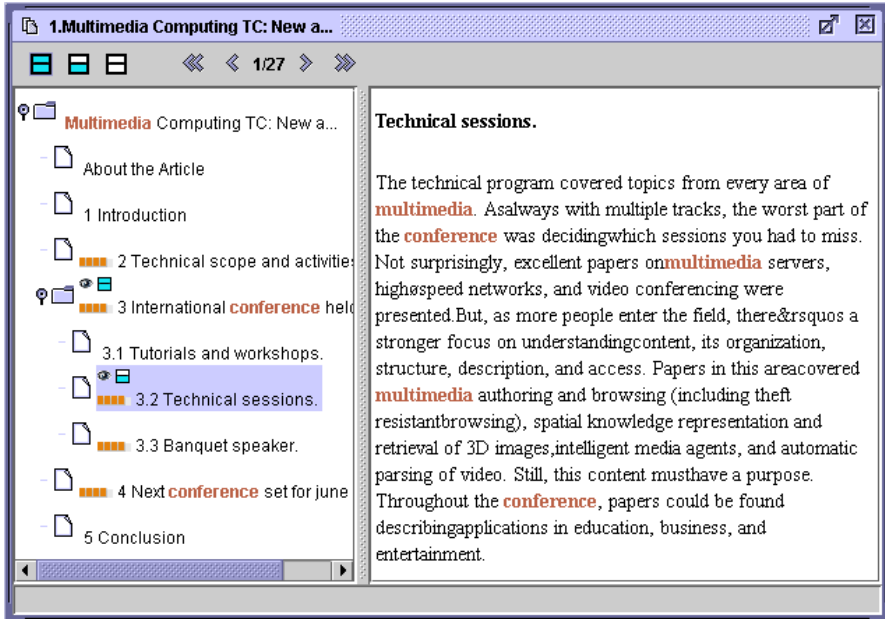


Fig. 2. Full text result in the baseline system used in Task A

4.5 Experimental Design

4.5.1 Experimental Matrix

A minimum of 6 test persons from each participating site were used. Each test person searched on one simulated work task from each category (chosen by the test person) as well as one of their own topics. The order in which task categories were performed by searchers was permuted in order to neutralise learning effects. This means that one complete round of the experiment requires 6 searchers.

The basic experimental matrix looked as follows:

- Rotation 1: OT, STG, STC
- Rotation 2: STC, OT, STG
- Rotation 3: STG, STC, OT
- Rotation 4: STG, OT, STC
- Rotation 5: STC, STG, OT
- Rotation 6: OT, STC, STG

Where OT = Own task, and STG, STC are the two 2 simulated work task categories. As can be seen from Table 1 above some groups did more than 6 test persons. It was attempted to coordinate the permutation rotations across these groups to arrive at an equal distribution of across the track.

4.5.2 Experimental Procedure

The experimental procedure for each test person is outlined below.

1. Experimenter briefed the searcher, and explained the format of the study
2. Tutorial of the system was given with a training task, and experimenter answered any questions
3. 'Instructions to searchers' handed out
4. Any questions answered by the experimenter
5. Entry questionnaire handed out
6. Task description for the first category handed out, and a task selected
7. Pre-task questionnaire handed out
8. Task began, and experimenter logged in. Max. duration 20 minutes. Experimenter logged out.
9. Post-task questionnaire handed out
10. Steps 7-10 were repeated for the two other tasks
11. Post-experiment questionnaire handed out
12. Interview

The system training, the three tasks and completion of questionnaires and interview were performed in one, continuous session. An 'Instructions to searchers' document gave information to the searchers about the experiment and their role in it, including basic information about system information, an outline of the experimental procedure, and how to assess elements for relevance. A number of questionnaires and guidelines for post-experiment interviews were provided by the track organisers. The purpose of the semi-structured interview was to attempt to elicit user perceptions of what they need from an element retrieval system, and to identify possible applications for element retrieval.

5 Task C

Task C was optional for those groups who wished to experiment with the Lonely Planet collection, and was separate from Task A and B. Thus additional test persons needed to be engaged for Task C. Task C was meant as an exploratory task to initiate interactive experiments with the LP collection.

5.1 Document Corpus

The document corpus used in Task C was the Lonely Planet collection. The Lonely Planet collection consists of 462 XML documents with information about destinations, which is particularly useful for travellers that want to find interesting details for their next holiday or business trip. The collection is called the "WorldGuide" and has been provided by the publishers of the Lonely Planet guidebooks. The collection not

only contains useful information about countries, but also includes information about interesting regions and major cities. For each destination an introduction is available, complemented with information about transport, culture, major events, facts, and an image gallery that gives an impression of the local scenery.

5.2 Relevance Assessments in Task C

A slightly more complex approach was taken for the collection of relevance assessments in Task C. The two-dimensional relevance scale was a modified version of a scale proposed at the INEX Methodology Workshop at the Glasgow IR Festival [6]. The relevance assessments were explained to the test persons as follows:

Two different dimensions are used to assess the relevance of an XML document component. The first determines the extent to which a document component *contains relevant information* for the search task. It can take one of the following three values: **highly relevant**, **somewhat relevant**, and **not relevant**. A document component is **highly relevant** if it covers aspects of the search task *without containing* too much non-relevant information. A document component is **somewhat relevant** if it covers aspects of the search task and at the same time *contains* much non-relevant information. A document component is **not relevant** if it *does not cover* any aspect of the search task.

The second relevance dimension determines the extent to which a document component *needs the context* of its containing XML document to make full sense as an answer. It can take one of the following three values: **just right**, **too large**, and **too small**. A document component is **just right** if it is reasonably *self-contained* and it needs *little* of the context of its containing XML document to make full sense as an answer. Alternatively, the document component can be either **too large** or **too small**. A document component is **too large** if it *does not need* the context of its containing XML document to make full sense as an answer. A document component is **too small** if it can only make full sense within the context of its containing XML document.

Given the above relevance values, the final assessment score of a document component can take one of the following **five** values:

- **Not Relevant (NR)** – if the document component does not cover any aspect of the search task;
- **Partial Answer (PA)** – if the document component is *somewhat relevant* (i.e. covers only some aspects of the search task) and *just right* (i.e. it is reasonably self-contained but still needs some of the context of its containing XML document to make full sense);
- **Exact Answer (EA)** – if the document component is *highly relevant* (i.e. covers all, or nearly all, aspects of the search task without containing too much non-relevant information) and *just right*;
- **Broad Answer (BA)** – if the document component is either *highly* or *somewhat relevant* and *too large* (i.e. it is reasonably self-contained and does not really need the context of its containing XML document to make full sense); and
- **Narrow Answer (NA)** - if the document component is either *highly* or *somewhat relevant* and *too small* (i.e. it is not self-contained and can only make full sense in the context of its containing XML document).

The test persons could select one of these values from a T-shaped relevance assessment box as shown in Fig. 4 and Fig. 5.

5.3 System

An interactive system for Task C was provided by Utrecht University. It is a fully functional element retrieval system which has been configured to suit Task C. There were two versions of the system: One which presented the results in context of the full text (i.e., highlighted), and an alternative version which presented the results in isolation. Fig. 3 shows the query ad result list interface common to both system versions. Fig. 4 and 5 shows the interface for the versions which showed results in context and isolated respectively.

5.4 Tasks/Topics

Eight topics that have previously been used for experiments with the Lonely Planet WorldGuide were selected and modified into short simulated work tasks for Task C. The tasks were arbitrarily split into 2 categories, and each test person searched two tasks from each category.

The screenshot shows a Mozilla Firefox browser window titled 'Structured Document Retrieval - Mozilla Firefox'. The address bar contains the URL <http://contentlab.cs.uu.nl/track/search/at69555a68125610515214763e09825f>. The page header includes the logo 'SD B³ R' and the text 'Universiteit Utrecht'. A search bar contains the query 'topical climate' and a 'Search' button. Below the search bar, it says 'Executed query: topical climate, against the lonelyplanet collection. Showing page: 1 of 10, with 50 results per page. Go to page: 2 3 4 5 6 7 8 9 10'. The results section is titled 'Results' and lists 14 items, each with a title and a URL. Annotations on the right side of the browser window point to three specific areas: 'Query field' (the search bar), 'Result list' (the list of search results), and 'Links to text' (the URLs in the search results).

Fig. 3. Query box and result list display used in Task C

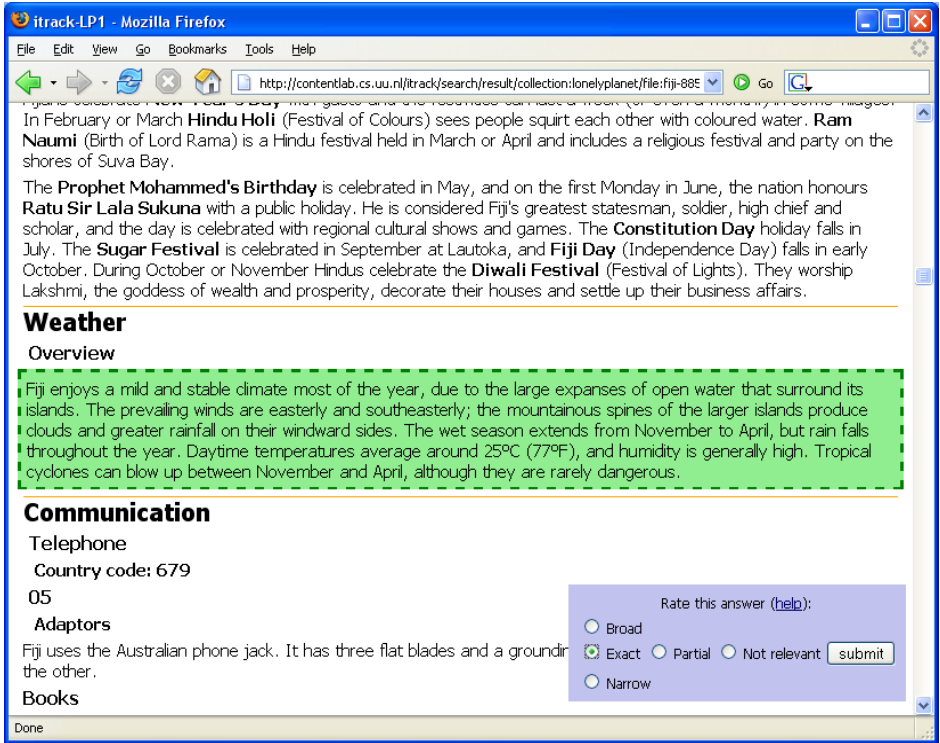


Fig. 4. Task C system version which presented the results highlighted in context of the full text

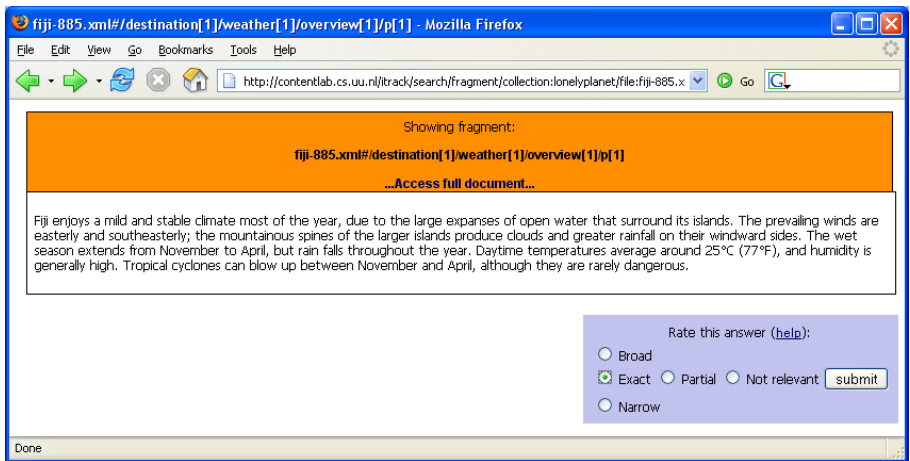


Fig. 5. Task C system version which presented the results in isolation

5.5 Experimental Design

A minimum of 4 test persons from each participating group were used in Task C. Each test person searched two simulated work tasks (chosen by the test person) from each of the two categories – a total of four per test person. The order in which task categories were performed was permuted in order to neutralise learning effects. This means that one complete round of the experiment required 4 searchers.

The basic experimental matrix looks as follows:

Rotation 1: Iso-C1, Cxt-C2

Rotation 2: Iso-C2, Cxt-C1

Rotation 3: Cxt-C1, Iso-C2

Rotation 4: Cxt-C2, Iso-C1

Where Iso = system with isolated results, and Cxt = system with results in context. C1 and C2 were the two simulated work task categories. The experimental procedure was very similar to the one used in Task A. However, no interview was conducted at the end of the experiment. A number of questionnaires were provided by the track organisers.

6 Concluding Remarks

In its second year, the Interactive Track at INEX looked into some fundamental questions surrounding interactive XML retrieval: does element retrieval make sense at all, do searchers prefer full-text to element retrieval, what applications could exist for interactive XML retrieval? In addition, the track also expanded by including two additional tasks and by attracting more participating groups. A total of 11 research groups and 119 test persons participated in the three different tasks that were included in the track.

In this paper, we have described the main issues that the Interactive Track at INEX 2005 attempts to address and the methodology and tasks that were used in the track. The data has now been released and it is the task of the participating groups to analyse and publish results from the track.

Acknowledgments

As track organisers we would like to thank the participating research groups for their work and input concerning the format of the track. In addition, we would like to thank Pia Borlund from the Royal School of Library and Information Science for help on constructing simulated work tasks, Claus-Peter Klas from University of Duisburg-Essen for help in running Daffodil, Roelof van Zwol, Sandor Spruit and Jeroen Baas, Utrecht University, for making their system available for Task C, and finally Jovan Pehceviski and co-workers for developing their relevance scale for use in the interactive track. Finally, we would like to thank Lonely Planets Publications Pty Ltd and the IEEE Computer Society for access to their data.

References

1. 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, Revised Selected Papers*. Berlin: Springer, p. 410-423. (Lecture Notes in Computer Science; 3493)
2. Larsen, B., Tombros, A., and Malik, S. (2004): *Interactive Track Workshop Report*. Slides presented at the INEX Workshop, December 2004 [<http://inex.is.informatik.uni-duisburg.de:2004/workshop.html#Reports>]
3. Borlund, P. (2003): The IIR evaluation model: a framework for evaluation of interactive information retrieval. In: *Information Research*, vol. 8, no. 3, paper no. 152. [Available at: <http://informationr.net/ir/8-3/paper152.html>]
4. Larsen, B., Tombros, A. and Malik, S. (2005): Obtrusiveness and relevance assessment in interactive XML IR experiments. In: Trotman, A., Lalmas, M. and Fuhr, N. eds. *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, held at the University of Glasgow, 30 July 2005. Second Edition*. Dunedin (New Zealand): Department of Computer Science, University of Otago, p. 39-42. [<http://www.cs.otago.ac.nz/inexmw/Proceedings.pdf>, visited 15-12-2005]
5. Pharo, N. and Nordlie, R. (2005): Context matters: an analysis of assessments of XML documents. In: Crestani, F. and Ruthven, I. eds. *Context: Nature, Impact, and Role : 5th International Conference on Conceptions of Library and Information Science, CoLIS 2005, Glasgow, UK, June 2005, Proceedings*. Berlin: Springer, p. 238-248. (Lecture Notes in Computer Science; 3507)
6. Pehcevski, J., Thom, J. A. and Vercoustre, A.-M. (2005): Users and assessors in the context of INEX: Are relevance dimensions relevant? In: Trotman, A., Lalmas, M. and Fuhr, N. eds. *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, held at the University of Glasgow, 30 July 2005. Second Edition*. Dunedin (New Zealand): Department of Computer Science, University of Otago, p. 47-62. [<http://www.cs.otago.ac.nz/inexmw/Proceedings.pdf>, visited 15-12-2005]
7. Kamps, J., de Rijke, M. and Sigurbjörnsson, B. (2005): What do Users Think of an XML Element Retrieval System?. In: *This volume*.
8. Klas, C.P., Albrechtsen, H., Fuhr, N., Hansen, P., Jacob, E., Kapidakis, S., Kovacs, L., Kriewel, S., Micsik, A., Papatheodorou, C., Tsakonas, G. (2006): *An Experimental Framework for Comparative Digital Library Evaluation: The Logging Scheme*. (Submitted for publication).