Seven years of INEX interactive retrieval experiments – lessons and challenges

Ragnar Nordlie and Nils Pharo

Oslo and Akershus University College of Applied Sciences
Postboks 4 St. Olavs plass, N-0130 Oslo, Norway
ragnar.nordlie@hioa.no; nils.pharo@hioa.no

Abstract. This paper summarizes a major effort in interactive search investigation, the INEX i-track, a collective effort run over a seven-year period. We present the experimental conditions, report some of the findings of the participating groups, and examine the challenges posed by this kind of collective experimental effort.

Keywords: User studies, interactive information retrieval, information search behavior

Published as: Nordlie, R. & Pharo, N. (2012). Seven years of INEX interactive retrieval experiments – lessons and challenges. In: Information access evaluation. Multilinguality, multimodality, and visual analytics. LNCS 7488 (pp. 13 – 23). Berlin Heidelberg: Springer-Verlag.

1 Introduction

The INEX interactive track was a run as a subtrack of the Initiative for the Evaluation of XML retrieval (INEX) every year from 2004 to 2010. In this track participating groups have followed a standard procedure for collecting data of end users performing search tasks in an experimental setting. This has made it possible to collect quite large data sets of user-system interaction under controlled conditions.

The INEX experiments started in 2002, when a collection of journal articles from IEEE was licensed for XML element retrieval experiments [1] to provide "an infrastructure to evaluate the effectiveness of content-oriented XML retrieval systems" [2]. The general assumption is that XML elements can be treated as candidate items for retrieval, similar to full text documents, document parts and document passages. The INEX experiments were designed following the TREC model, with a test collection consisting of documents, topics/tasks (submitted by the participating groups), and relevance assessments provided by the participants, thus making it possible to compute the retrieval effectiveness of different matching algorithms. Since its beginning several tracks have been introduced to the initiative in order to explore topics such as relevance feedback, heterogeneous collections, natural queries, document mining,

multimedia; and also a track devoted to studying interactive information retrieval of XML-coded data through user experiments.
In this paper we will discuss some of the lessons learnt throughout the seven years of these interactive experiments. We start by presenting the experimental conditions of the interactive track (hereafter the i-track). Then we will explore some of the findings made during the years. In the third part we will discuss the possible levels of interpretation for INEX i-track data and finally we will point out some of the challenges and problems we have experienced.

2 INEX i-track experimental conditions

The design of the i-track user experiments has followed rather similar patterns throughout the years. The elements used are:

- A search system developed by the track organizers. Optionally participants in the track developed their own system for additional experiments
- A document corpus, often the same that was used as the test collection for the standard ad hoc-track
- A set of topics or simulated tasks to be searched for by the experiment subjects
- Questionnaires, either paper based or integrated in the online experimental setup
- A relevance assessment scale for relevance assignments by searchers
- A system for recording transaction logs
- A standard procedure for data collection

We shall look at the details of each of these items.

2.1 The search system

Since its beginning the i-track organizers have made available a search system for the participating groups to use. The system used in 2004 [3] was based on the HyREX retrieval engine [4]. The system was designed for XML retrieval and when queried returned a ranked list of XML elements, where each element was accompanied with the title and author of the source document of the element, its retrieval value, and its XPath. In 2005 [5] the organizers switched to a system built within the Daffodil retrieval framework [6], which provided some improvements over the previous system, specifically with respect to handling of overlapping elements, improved element summaries, and supportive interface functionalities. The Daffodil system was also used in 2006 [7], but this year in two different versions; one using a passage retrieval backend and the other an element retrieval backend. In 2008 [8] and 2009 [9] the element retrieval version of Daffodil was also used. In 2010 [10] a new system was

developed based on the ezDL framework[1], which resides on a server and is maintained by the University of Duisburg-Essen.

2.2 The document corpora

In total three different document collections have been used in the i-track. In 2004 and 2005 a collection of computer science journal articles published by IEEE was made available for the experiments. The same collections were used by the INEX ad hoc-track, with additional documents added in the 2005 collection (see Table 1).
In 2006 and 2008 the Wikipedia collection [11] was used, it consists of more than 650 000 articles collected from the English version of Wikipedia. The last two years (2009-2010) the Amazon/LibraryThing corpus was put together for the i-track: "[t]he collection contains metadata of 2 780 300 English-language books. The data has been crawled from the online bookstore of *Amazon* and the social cataloging web site *LibraryThing* in February/March 2009 by the University of Duisburg-Essen. The MySQL database containing the crawled data has a size of about 190 GB. Cover images are available for over one million books (100 GB of the database). Several millions of customer reviews were crawled" [10]. This collection is currently also in use by the INEX book track.

Table 1. Document corpora used in the i-track

Year	Collection	Size (no of items)	Use
2004-2005	IEE journals	12107/16819	Ad hoc & i-track
2006-2008	Wikipedia articles	659 388	Ad hoc & i-track
2009-2010	Amazon/Librarything	2 780 300	i-track

2.3 Topics and tasks

The topics or tasks used in the i-track experiments were developed for exploring a variety of research questions. Borlund's [12] simulated work task methodology was used to formulate the tasks in order to make it clearer for the searcher which type of context the task intended to represent. In Table 2 we see a summary of the task categories and the number of tasks to be performed by the searchers.
In 2004 a selection of content only (CO) topics from the ad hoc-track was selected. The topics were picked to represent two different categories of tasks, "background tasks" (B) and "comparison tasks" (C) [3]. The selection of categories was justified from studies that have shown that different types of tasks invoke different relevance criteria for assessing web pages [13]. It turned out that the 2004 categorization was not a "great success" therefore in 2005 task categories were simplified to "general" (G) and "challenging" (C) tasks [5] and tasks representing these categories were col-

[1] http://www.ezdl.de/

lected from the ad hoc-tasks. In addition, the searchers in the 2005 i-track were asked to formulate examples of their own information needs to be used as "own" tasks. In 2006, using the Wikipedia collection, the organizers wished to emphasize the effect of different task types and created "a multi-faceted set of twelve tasks […] with three task types" [7]: "decision making", "fact finding", and "information gathering". These were, in turn, split into two structural kinds ("Hierarchical" and "Parallel"). The selection of task categories was based on work done by Elaine Toms and her colleagues [14]. In 2008 a new set of tasks were used, "intended to represent information needs believed to be typical for Wikipedia users" [8], the two categories were "fact-finding tasks" and "research tasks". With the Amazon/LibraryThing collection new task sets were introduced, in 2009 the searchers were asked to formulate a task on their own given the premises that they should find a textbook within a course they were attending. In addition two task categories were developed by organizers, "broad tasks" which "were designed to investigate thematic exploration" and "narrow tasks" representing "narrow topical queries" [9]. A similar design of tasks were used in 2010 [10], but the categories were now called "explorative" and "data gathering".

Table 2. Tasks used in the i-track

Year	Task categories	Tasks per category	Tasks per searcher
2004	Background; Comparison	2	2
2005	General; Challenging; Own	3 (+ own)	3
2006	Decision making; Fact finding; Information gathering	4 (2 of each structure)	4
2008	Fact-finding; Research	3	2
2009	Broad; Narrow; Own	3 (+ own)	3
2010	Explorative; Data gathering; Own	3 (+ own)	3

2.4 Questionnaires

The questionnaires distributed in the i-track experiments have not changed a lot during the years. Experiment participants have been asked to answer the following types of questionnaires:

1. A pre-experiment questionnaire with questions about the participants' background, including demographic questions, education, search experience and experience with different types of information sources
2. Pre-task questionnaires with questions about the participants' task familiarity and the perceived difficulty of the task
3. Post-task questionnaires on the experienced task difficulty and perceived satisfaction as well as on system features related to the task
4. Post-experiment questionnaires on general system related issues

2.5 Relevance assessments scales

The recognition of relevance as a more subtle and dynamic feature in IR has led to the introduction of non-binary relevance assessments in IR system evaluation [15]. In the i-track experiments many different relevance scales have been used to try to learn about the relationship between elements, their context and how end users react to the levels of granularity explicated in XML retrieval systems.
In the 2004 i-tack experiments a two-dimensional relevance scale was used, it was designed to measure how "useful" and how "specific" the assessed element was in relation to the search task [3]. Each dimension had three degrees of relevance which (with the additional value of "not relevant") made a total of 10 possible dimensions (see Table 3.

Table 3. The INEX 2004 i-track relevance scale

Value	Explanation
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 & Very specific
H	Marginally useful & Fairly specific
I	Marginally useful & Marginally specific
J	Contains no relevant information

In 2005, 2009 and 2010 organizers used a three level relevance scale, asking searchers to state if the elements were "relevant", "partially relevant" and "not relevant". In 2006 and 2008 a two-dimensional scale was also used, although a bit different from the 2004-scale. This scale was based on the work of Pehcevski [16] and aimed to balance the need for information on the perceived granularity of retrieved elements and their degree of relevance, and was intended to be simple and easy to visualize [7]. Figure 1 shows how the system interface presented the relevance scale to the searchers.

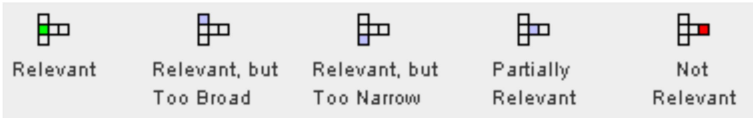

Fig. 1. INEX 2006 and 2008 interactive track relevance assessments scale

2.6 System transaction logs

For each of the experiments transaction logs have been recorded by the systems. These logs capture all events during searchers' system-interaction, including their use of search facilities, their queries, the query results, all elements viewed, and all relevance assessments made. The logs have been recorded as XML-files. In addition, some participating institutions have at different times used more sophisticated recording devices, such as screen capture programs to track mouse movements, or eye-tracking devices.

2.7 Data collection procedures

Data collection has followed a very similar procedure from each year to the next, the following procedure is quoted from [8]:

1. Experimenter briefed the searcher, and explained the format of the 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 were answered by the experimenter.
4. The control system administered the pre-experiment questionnaire.
5. Topic descriptions for the first task category was administered by the system, and a topic selected
6. Pre-task questionnaire was 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. The task ended by clicking the "Finish task" button.
8. Post-task questionnaire was administered.
9. Steps 5-8 were repeated for the second task.
10. Post-experiment questionnaire was administered

3 INEX i-track findings

Analysis of INEX i-track data has been reported in the annual INEX proceedings and in the SIGIR Forum, at conferences such as SIGIR [17] and IIIX [18] and in scientific journals, for example Information Processing & Management [19] and JASIST [20].
In principle, the data collected in the INEX experiments allow for interpretation on at least three different levels. The focus might be on the *types* of transactions / actions over the whole collection of searches, without regard to individual searchers or individual sessions. This represents a very quantitative view of search behavior, and includes investigations of how many times a text element on a certain level of granularity is viewed, judged relevant, with which degree of confidence, at what stage in the search etc.

Alternatively, the focus might be on *patterns* of transactions, again over the whole collection of searches. This approach attempts to answer questions such as what sequences of document or text element views precedes a relevance decision, how queries are developed and what influence factors such as the documents viewed in the search process has on query development, or where in the session a certain behavioral pattern occurs.

The third level of investigation would look at individual *sessions*, or sequences of interactions within sessions, to try to understand how factors such as user characteristics or types of search purpose influence actions, transaction patterns, or relevance decisions. On this level, quantification would be subordinate to a more qualitatively based analysis.

The research based on i-track data has in particular, not very surprisingly, focused on the element types users prefer to see when interacting with XML retrieval systems [17–23]. For the most part, this research has been based on the *type of transaction* perspective described above, and has examined the total corpus of search sessions as a set of countable instances of element views and associated relevance decisions. At times, these transaction counts have been subdivided by factors such as the task type or the systems' presentation format, but the perspective has still been to isolate the single transaction occurrences and quantify results.

Data from the 2004 i-track is analyzed in [21], the authors found that section elements were judged to be the most relevant element both with respect to specificity and to usefulness. In cases when both full articles and sections of the same article were assessed the articles were often assessed as more relevant than their section elements.

2005 i-track data are analyzed in [17, 19], which report that most users first accessed the "front matter" element (mainly containing metadata) when examining a document, but it is speculated that this should be interpreted as users wishing to obtain the full article first. In [23] the influence of topic knowledge, task type and user motivation on users element type preferences is analyzed. The authors find that users' topic familiarity is an important factor in estimating the type of task s/he is performing. [22] compared i-track 2004 and 2005 data with respect to how two different interfaces for presentation of query results (unstructured and hierarchically structured) impacts element assessments. The authors found that there was a stronger tendency for searchers to assess section elements, compared to other elements, when elements from the same document were scattered in the result list instead of presented structurally under the full article. [18] performed an analysis of interaction with the 2005 i-track Lonely planet collection, and found that that the major part of "exact" relevance assessments were made on elements at a more fine-grained level of granularity than the full document.

2006 i-track data was analyzed in [20], where it was found that larger units of text such as full articles and sub-sections were considered of most use for the searchers. The tendency was stronger for searches involving information-gathering tasks.

4 INEX i-track as model for interaction studies

There are obvious advantages to attempting the kind of collective, decentralized, semi-controlled experiment which the INEX interactive effort represents. It is possible, at least in theory, to collect a number of search sessions for analysis which would be extremely time-consuming for each institution to acquire on its own, and which, again in theory, should make it possible to draw quantifiable, not only qualitative conclusions. The data should be possible to compare across years, and be available for analysis by other than the initial experimenters. The relatively rich background data on the participating searchers should allow for quite detailed interpretation of the data. On the other hand the decentralized data collection makes a controlled selection of searchers impossible, so that the sample will be self-selected. Even if the main research objectives are shared by the participants, the pooling of data also makes it difficult to have firmly stated research questions and thus establish and maintain the necessary control of the variables influencing the search activities under study. The research based on the INEX data has revealed a number of problems and challenges which need to be addressed in future interaction studies of this kind.

4.1 Tasks, data and systems

The *tasks* assigned to the searchers have attempted to emulate search situations which might conceivably call for different search behaviors and search result contents. The variation in tasks over the years shows the difficulty of finding a good theoretical fundament to base these distinctions on. It also makes it difficult to compare results across years. The challenge has been to find tasks that at the same time match real-life search situations, are uniformly understandable without specialist knowledge, are not prone to too much individual interpretation, and are sufficiently challenging to engage the searchers on whom the tasks are imposed. When searchers are given a selection of tasks intended to represent the same search situation, it is particularly important that these conditions are satisfied. In actual fact, even if searchers have been asked about level of task familiarity it has been difficult to control for differences in interpretation and level of involvement. In the years when a self-selected task has been included, it has been particularly difficult to specify this in a way which allows meaningful interpretation and comparison.

The choice of *database* has attempted to represent a set of data that is at the same time realistic and controllable and provides interpretable results. Again, the difference between the three data sets used makes comparison between years difficult. Relevance judgment is a very different task when applied to articles or parts of articles in a heavily technical domain as represented by the IEEE corpus, as opposed to relatively brief, well-structured and popularized Wikipedia articles, and judging relevance when the full text is available is again a different task from judging the relevance of books when only metadata are available, no matter how extensively the metadata represent them.

The concept of *relevance* in itself constitutes a challenge. The large variation in measures of relevance applied in the i-track over the years illustrates the difficulty of establishing a metric which is both understandable and applicable by the searchers, and which at the same time measures with sufficient precision the success or failure of the behaviors or the system features under investigation. Since the main purpose of INEX has been to investigate the effects of the facility to present elements of text of different granularity to searchers, it has been important to measure some kind of degree of relevance related to the level of granularity presented. At the same time there is evidence that searchers are not able to interpret and apply a complex relevance measure consistently, and it is also difficult to determine which of the features of the complex measure to take into consideration when analyzing the interactions.
The *search system* has also varied over the course of the experiments. For the most part, searchers have been exposed to a system which they have not had the opportunity to use previously. This has the advantage of eliminating possible effects of system familiarity, but under the time constraints posed by the laboratory conditions of the experiments, it has been difficult to ensure a common understanding of system functionalities in the training time available, and the decentralized data collection further complicates a common presentation of the system or, in some years, systems. It has proven difficult to identify and isolate the effect of different degrees of mastery of the system as distinct from different search styles or different understanding of the tasks.

4.2 Units and levels of analysis

The abovementioned problem illustrates a major challenge with interaction studies in general and particularly with the i-track experiments: how is it possible to identify and isolate the features (of users, interfaces, tasks…) which may influence or explain behavior? Is it task variations, different understandings of the interface, different level of training, different level of interest in the experiment, differences in search experience, age or education, or other factors, which prompts certain actions to be taken or features to be used? To a certain extent, the responses to the questionnaires may clarify this, but the complex interrelationship between the factors is difficult to capture. This becomes particularly problematic when much of the interpretation of the data, as mentioned earlier, is based on counts of transactions or actions rather than on analysis of sessions.
A major challenge with the interpretation of the i-track data is the identification and specification of what constitutes a unit of analysis. In the logs, it is possible to identify individual actions, such as browsing a list of references, choice of an article or a smaller unit of text to view, etc. It is also possible to see elapsed time between actions. It is of course also possible to interpret these actions as parts of a sequence constituting a transaction, such as the series of browse and view actions which precede a relevance judgment. The difficulty is both to decide what sequences of actions should be considered part of a meaningful transaction and which are random sequences, how to delimit and define the transactions and how to agree on what constitutes a meaningful transaction. Also, there are actions or occurrences which are important for un-

derstanding search behavior and which are impossible or difficult to determine on the basis of search logs, such as reading behavior, handling of disruptions etc. Techniques for capturing such data have been attempted within the i-track framework, such as eye tracking, screen capture, thin-aloud protocols etc, but such data are not easily shareable, and they open new interpretational challenges of their own.

It has proven difficult to use the i-track studies to determine the usefulness of XML coding of text to support users' search. This is both because of the difficulty of interpreting the data with any degree of certainty, as discussed above, and because the concept of XML search itself is poorly defined – it is for instance difficult to distinguish a system based on XML coding from a passage retrieval system from a user point of view, at least as long as semantic XML coding is still difficult to attain and exploit.

With all these constraints and their problematic features, however, the i-track data still constitute a rich source of interaction data which still only has been tapped to a certain extent. More importantly, the i-track data and the i-track experience might conceivably form the basis of the development of a framework or frameworks for user search investigation which may supply more firmly described and shareable data than those we have discussed here,

5 References

1. Gövert, N., Kazai, G.: Overview of the Initiative for the Evaluation of XML retrieval (INEX) 2002. Presented at the INEX Workshop (2002).
2. Kazai, G., Lalmas, M., Fuhr, N., Gövert, N.: A report on the first year of the INitiative for the evaluation of XML retrieval (INEX'02). Journal of the American Society for Information Science and Technology. 55, 551–556 (2004).
3. Tombros, A., Larsen, B., Malik, S.: The interactive track at INEX 2004. In: Fuhr, N., Lalmas, M., Malik, S., and Szlávik, Z. (eds.) Advances in XML Information Retrieval. pp. 410–423. Springer, Berlin (2005).
4. Fuhr, N., Gövert, N., Großjohann, K.: HyREX: Hyper-media Retrieval Engine for XML. University of Dortmund, Computer Science (2002).
5. Larsen, B., Malik, S., Tombros, A.: The interactive track at INEX 2005. In: Fuhr, N., Lalmas, M., Malik, S., and Kazai, G. (eds.) Advances in XML Information Retrieval and Evaluation. pp. 398–410. Springer, Berlin (2006).
6. Gövert, N., Fuhr, N., Klas, C.-P.: Daffodil: Distributed Agents for User-Friendly Access of Digital Libraries. Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries. pp. 352–355. Springer, Berlin (2000).
7. Malik, S., Tombros, A., Larsen, B.: The Interactive Track at INEX 2006. In: Fuhr, N., Lalmas, M., and Trotman, A. (eds.) Comparative Evaluation of XML Information Retrieval Systems. pp. 387-399. Springer, Berlin (2007).
8. Pharo, N., Nordlie, R., Fachry, K.N.: Overview of the INEX 2008 Interactive Track. In: Geva, S., Kamps, J., and Trotman, A. (eds.) Advances in Focused Retrieval. pp. 300-313. Springer, Berlin (2009).
9. Pharo, N., Nordlie, R., Fuhr, N., Beckers, T., Fachry, K.N.: Overview of the INEX 2009 Interactive Track. In: Geva, S., Kamps, J., and Trotman, A. (eds.) Focused Retrieval and Evaluation. pp. 303-311. Springer, Berlin (2010).

10. Pharo, N., Beckers, T., Nordlie, R., Fuhr, N.: Overview of the INEX 2010 Interactive Track. In: Geva, S., Kamps, J., Schenkel, R., and Trotman, A. (eds.) Comparative Evaluation of Focused Retrieval. pp. 227-235. Springer, Berlin (2011).
11. Denoyer, L., Gallinari, P.: The Wikipedia XML corpus. SIGIR Forum. 40, 64–69 (2006).
12. Borlund, P.: Evaluation of interactive information retrieval systems. Abo Akademis Forlag (2000).
13. Tombros, A., Ruthven, I., Jose, J.M.: How users assess Web pages for information seeking. Journal of the American Society for Information Science and Technology. 56, 327-344 (2005).
14. Toms, E.G., O'Brien, H., Mackenzie, T., Jordan, C., Freund, L., Toze, S., Dawe, E., Macnutt, A.: Task effects on interactive search: the query factor. In: Fuhr, N., Kamps, J., Lalmas, M., and Trotman, A. (eds.) Focused Access to XML Documents. pp. 359–372. Springer, Berlin (2008).
15. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '00. pp. 41-48. , Athens, Greece (2000).
16. Pehcevski, J.: 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. University of Otago, Dunedin (2006).
17. Larsen, B., Tombros, A., Malik, S.: Is XML retrieval meaningful to users?: searcher preferences for full documents vs. elements. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 663–664. ACM, New York, NY, USA (2006).
18. Hammer-Aebi, B., Christensen, K.W., Lund, H., Larsen, B.: Users, structured documents and overlap: interactive searching of elements and the influence of context on search behaviour. Proceedings of the 1st international conference on Information interaction in context. pp. 46–55. ACM, New York, NY, USA (2006).
19. Pharo, N.: The effect of granularity and order in XML element retrieval. Information Processing and Management. 44, 1732–1740 (2008).
20. Pharo, N., Krahn, A.: The effect of task type on preferred element types in an XML-based retrieval system. Journal of the American Society for Information Science and Technology. 62, 1717-1726 (2011).
21. Pharo, N., Nordlie, R.: Context Matters: An Analysis of Assessments of XML Documents. In: Crestani, F. and Ruthven, I. (eds.) Context: Nature, Impact, and Role. pp. 1911-1912. Springer Berlin (2005).
22. Kim, H., Son, H.: Users Interaction with the Hierarchically Structured Presentation in XML Document Retrieval. In: Fuhr, N., Lalmas, M., Malik, S., and Kazai, G. (eds.) Advances in XML Information Retrieval and Evaluation. pp. 422-431. Springer Berlin (2006).
23. Ramírez, G., de Vries, A.P.: Relevant contextual features in XML retrieval. Proceedings of the 1st international conference on Information interaction in context. pp. 56–65. ACM, New York, NY, USA (2006).