Report on INEX 2010

D. Alexander P. Arvola T. Beckers P. Bellot T. Chappell C.M. De Vries
A. Doucet N. Fuhr S. Geva J. Kamps G. Kazai M. Koolen
S. Kutty M. Landoni V. Moriceau R. Nayak R. Nordlie N. Pharo
E. SanJuan R. Schenkel A. Tagarelli X. Tannier J.A. Thom A. Trotman
J. Vainio Q. Wang C. Wu

Abstract

INEX investigates focused retrieval from structured documents by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results. This paper reports on the INEX 2010 evaluation campaign, which consisted of a wide range of tracks: Ad Hoc, Book, Data Centric, Interactive, QA, Link the Wiki, Relevance Feedback, Web Service Discovery and XML Mining.

1 Introduction

Traditional IR focuses on pure text retrieval over "bags of words" but the use of structure—such as document structure, semantic metadata, entities, or genre/topical structure—is of increasing importance on the Web and in professional search. INEX has been pioneering the use of structure for focused retrieval since 2002, by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results.

Focused retrieval takes many forms. Hence, the INEX 2010 evaluation campaign consisted of a wide range of tracks:

Ad Hoc Track The main track of INEX 2010 will be investigating the effectiveness of XML-IR and passage retrieval for highly focused retrieval by restricting result length to "snippets" or discounting for reading effort, using Wikipedia as a corpus.

Book Track Investigating techniques to support users in reading, searching, and navigating the full texts of digitized books, by constructing reading lists of books for a given topic, or by looking for book pages that refute or confirm a factual statement.

Data Centric Track Investigating focused retrieval over a strongly structured collection of IMDb documents, containing information about various entities like movies, actors, directors.

Interactive Track Investigating the behavior of users when interacting with XML documents, as well as developing retrieval approaches which are effective in user-based environments, working on a Amazon corpus combining formal book descriptions and user-generated data.

Link the Wiki Track Investigating link discovery in the Te Ara encyclopedia.

Question Answering Track Investigating real-world focused information needs formulated as natural language questions, using the collection structure to construct readable summaries of question context, and lists of answers.

Relevance Feedback Track Investigate the utility of incremental passage level feedback by simulating a searcher's interaction, with submissions in the form of a executable computer program rather than a list of search result.

Web Service Discovery Investigate techniques for discovery of Web services based on searching service descriptions provided in WSDL.

XML Mining Track Investigating structured document mining, especially the classification and clustering of semi-structured documents.

In the rest of this paper, we discuss the aims and results of the INEX 2010 tracks in relatively self-contained sections: the Ad Hoc track (Section 2), the Book track (Section 3), the Data Centric track (Section 4), the Interactive track (Section 5), the QA track (Section 6), the Link the Wiki track (Section 7), the Relevance Feedback track (Section 8), the Web Service Discovery track (Section 9), and the XML Mining track (Section 10).

2 Ad Hoc Track

In this section, we will briefly discuss the aims of the INEX 2010 Ad Hoc Track, its tasks and setup, the used measures and results. Further details are in [1].

2.1 Aims and Tasks

The main novelty of the Ad Hoc Track at INEX 2010 is its focus on retrieval under resource restricted conditions such as a small screen mobile device or a document summary on a hit-list. Here, retrieving full articles is no option, and we need to find the best elements or passages that convey the relevant information in the Wikipedia pages. So one can view the retrieved elements/passages as extensive result snippets, or as an on-the-fly document summary, that allow searchers to directly jump to the relevant document parts. This leads to variants of the focused retrieval tasks that address the impact of result length or reading effort, thinking of focused retrieval as a form of "snippet" retrieval.

The INEX 2010 Ad Hoc Track featured three variants of ad hoc retrieval against a collection of structured documents, factoring in various restrictions on the length of retrieved results (Relevant in Context, Restricted Relevant in Context, and Restrict Focused). In addition, there was an Efficiency Task enabling a systematic study of efficiency-effectiveness trade-offs with the different systems. In order to extend the ad hoc retrieval test collection on the INEX 2009 Wikipedia Collection with additional topics and judgments, the Ad Hoc track topics and assessments stayed unchanged. The change of task calls for new measures that factor in reading effort [2], which are compared to the earlier INEX measures based on passage level precision and recall [7].

2.2 Test Collection

The Ad Hoc Track uses a document collection based on the English Wikipedia. The original Wiki syntax has been converted into XML, using both general tags of the layout structure

(like *article*, *section*, *paragraph*, *title*, *list* and *item*), typographical tags (like *bold*, *emphatic*), and frequently occurring link-tags. The annotation is enhanced with semantic markup of articles and outgoing links, based on the semantic knowledge base YAGO, explicitly labeling more than 5,800 classes of entities like persons, movies, cities, and many more [11].

INEX has been pioneering peer-topic creation and peer-assessments since 2002. At INEX 2009, a total of 115 ad hoc search topics were created by participants. The topics were assessed by participants following precise instructions. The assessors used the GPXrai assessment system that assists assessors in highlighting relevant text. Topic assessors were asked to mark all, and only, relevant text in a pool of 750 documents. After assessing an article with relevance, a separate best entry point decision was made by the assessor. The relevance judgments were frozen on November 3, 2010. At this time 52 topics had been fully assessed. Moreover, for 7 topics the judgments of a second judge are also available.

2.3 Results

We received a total of 213 submissions from 18 participating groups, which are discussed in detail in the papers in [5].

The first goal was to study focused retrieval under resource restricted conditions such as a small screen mobile device or a document summary on a hit-list. That is, to think of focused retrieval as a form of "snippet" retrieval, suggesting either by measures that factor in reading effort or by tasks that have restrictions on the length of results. The results of the effort based measures are a welcome addition to the earlier recall/precision measures. It addresses the counter-intuitive effectiveness of article-level retrieval—given that ensuring good recall is much easier than ensuring good precision. As a result there are significant shifts in the effectiveness of systems that attempt to pinpoint the exact relevant text, and are effective enough at it. Having said that, even here locating the right articles remains a prerequisite for obtaining good performance.

The second goal was to examine the trade-off between effectiveness and efficiency. Participants were asked to report efficiency-oriented statistics for their Ad Hoc-style runs on the 2010 Ad Hoc topics. The average running times per topic varied from 1ms to 1.5 seconds, where the fastest runs where run on indexes kept in memory. This is almost an order of magnitude faster than the fastest system from INEX 2009, and the low absolute response times demonstrate that the current collection is not large enough to be a true challenge. Result quality was comparable to other runs submitted to other tasks in the Ad Hoc Track.

2.4 Outlook

This was the fifth year that INEX has studied ad hoc retrieval against the Wikipedia. In 2006–2008 the English Wikipedia of early 2006 transformed into XML was used (containing 659,338 Wikipedia articles), which was updated to version of late 2008 (containing 2,666,190 Wikipedia articles and incorporating semantic annotations from YAGO) and used in 2009–2010. The test collections on Wikipedia have large sets of topics, 291 for the 2006–2008 Wikipedia and 120 for the 2009–2010 Wikipedia. There are relevance judgments at the passage level (both best-entry-points as well as the exact relevant text) plus derived article-level judgments resulting in an attractive document retrieval test collection using freely available documents in a non-news genre. There is a range of evaluation measures for evaluating the various retrieval tasks [7, 2], in addition to the standard measures that can be used for

article-level retrieval. Moreover, there is rich information on topic authors and assessors, and their topics and judgments based on extensive questionnaires, allowing for detailed further analysis and reusing topics that satisfy particular conditions.

After five years, there seems little additional benefit in continuing with focused retrieval against the Wikipedia corpus, given that the available test collections that are reusable in various ways. It is time for a new challenge, and other tracks have started already addressing other aspects of ad hoc retrieval: the INEX 2010 Book Track using a corpus of scanned books, the INEX 2010 Data Centric Track using a corpus of IMDb data, and the INEX 2010 Interactive Track using a corpus of Amazon and Library Thing data.

3 Book Track

In this section, we will briefly discuss the aims and tasks of the INEX 2010 Book Track, the test collection, and the results. Further details are in [8].

3.1 Aims and Tasks

The goal of the INEX Book Track is to evaluate approaches for supporting users in searching, navigating and reading the full texts of digitized books. In 2010, the investigation focused around four tasks:

- The Best Books to Reference (BB) task, framed within the user task of building a reading list for a given topic of interest, aims at comparing traditional document retrieval methods with domain-specific techniques, exploiting book-specific features, e.g., back-of-book index, or associated metadata, e.g., library catalogue information;
- The Prove It (PI) task aims to test focused retrieval approaches on collections of books, where users expect to be pointed directly at relevant book parts that may help to confirm or refute a factual claim;
- The Structure Extraction (SE) task aims at evaluating automatic techniques for deriving structure from OCR data and building hyperlinked table of contents;
- The Active Reading task (ART) aims to explore suitable user interfaces to read, annotate, review, and summarize multiple books.

A total of 93 organizations registered for the track, but only 12 groups took an active role. In this section, we will briefly discuss the BB and PI tasks. Further details on all four tasks are available in [8].

3.2 Test Collection

The Book Track builds on a collection of over 50,000 out-of-copyright books of different genre (e.g., history books, text books, reference works, novels and poetry) marked up in XML.

A total of 83 new topics were contributed to the test collection in 2010, both by INEX participants and by workers on Amazon's Mechanical Turk (AMT) service, a popular crowdsourcing platform.

Similarly to the topics, relevance assessments were collected from INEX participants as well as crowdsourced through AMT. The INEX judgments were used as gold set in the relevance gathering task on AMT for a selected set of 21 (out of the 83) topics. For each

topic, a separate Human Intelligence Task (HIT) was published with the title of the HITs reflecting the subject area of the topic in order to attract workers with interest in the subject. Each HIT consisted of 10 pages to judge, where at least one page was already labeled as relevant by an INEX participant, enabling us to check the quality of the worker's work. In each batch of HITs, we published 10 HITs per topic and thus collected labels for 100 pages per topic from 3 workers, obtaining a total of 6,300 labels. When constructing the AMT assessment pools, we combined three different pooling strategies: top-n, rank-boosted, and answer boosted, with the aim to get i) a good coverage of the top results of the official PI runs, ii) a large overlap with the pages judged by INEX assessors, and iii) to maximize the number of possibly relevant pages in the pool to improve reusability. As a result of the mixed pooling methods, in each 100 page assessment pool we have roughly the top 30 pages per pooling method plus the known relevant pages. Pages can occur only once in each HIT, but the known relevant pages could occur in multiple HITs, leading to 1,918 query/page pairs.

For the 21 topics, a total of 5,975 page-level relevance labels were collected from 7 INEX participants and 6,300 labels from 194 workers on AMT (note: this excludes 145 HITs by 2 workers that were rejected). The AMT set contains 3 judgments per page, while the INEX data contains only one label per page (apart from 430 pages that were assessed multiple times, with 91% agreement). An analysis of the AMT labels showed high agreement with the INEX labels: 71% agreement based on four levels of relevance degree and 78% for binary relevance. The consensus among AMT labels is 92% for binary relevance (90% for the four levels), meaning that on average the majority vote for a label forms 92% of all workers votes.

From the multiple labels per page, we derived a single judgment for evaluation, using majority vote among the AMT labels and merging with the INEX set by selecting an INEX label over an AMT label for the same page. The resulting set contains 6,527 page level judgments, including 489 pages that confirm or refute the factual claim of a topic (23 per topic on average) and 719 pages that are relevant to the topic of the factual statement (34 per topic).

In addition to page level judgments, 990 book level judgments were collected for the 21 topics from the task organizers for the evaluation of the BB task.

3.3 Results

A total of 15 BB runs were submitted by 3 groups (2 runs by University of Amsterdam; 4 runs by University of California, Berkeley; and 9 runs by the University of Avignon), and a total of 10 PI runs by 3 groups (4 runs by the University of Amsterdam; 5 runs by Oslo University College; and 1 run by the University of Avignon).

The best BB run (NDCG@10=0.6579) was submitted by the University of California, Berkeley (p14-BOOKS2010_T2_PAGE_SUM_300) who employed page level retrieval methods and derived book level scores by summing the page level scores within the books. Page level scores were generated using a probabilistic approach based on logistic regression. A run by the University of Avignon followed close second with NDCG@10=0.6500. They experimented with a method for correcting hyphenations in the books and used the language modeling approach of the Lemur toolkit.

The best PI run (NDCG@10=0.2946) was submitted by the University of Amsterdam (p6-inex10.page.fb.10.50), who investigated the impact of varying the units of retrieval, e.g., books, individual pages, and multiple pages as units in the PI task. They achieved best performance with their individual page level index and using pseudo relevance feedback.

3.4 Outlook

In 2010, we piloted a test collection construction method crowdsourcing both topics and relevance labels. With our quality control rich HIT template, we obtained high quality labels showing 78% agreement with INEX gold set data. This has paved the way to completely removing the burden of relevance assessments from the participants in 2011.

In 2011, the track will shift focus onto more social and semantic search scenarios, while also continuing with the AR and SE tasks. The track will build on its current book corpus as well as a new collection from Amazon Books and LibraryThing.com. The PI task will run with minor changes, also asking systems to differentiate positive and negative evidence for a given factual claim. The BB task will be replaced by the new Social Search for Best Books (SSBB) task which will build on the corpus of 1.5 million records from Amazon Books and LibraryThing.com. SSBB will investigate the value of user-generated metadata, such as reviews and tags, in addition to publisher-supplied and library catalogue metadata, to aid retrieval systems in finding the best, most relevant books for a set of topics of interest.

4 Data Centric Track

In this section, we will briefly discuss the aims, data, tasks, and results of the INEX 2010 Data Centric Track. Further details are in [14].

4.1 Aims

At INEX 2009 Qiuyue Wang suggested that INEX should reexamine the question of information retrieval search over highly structured search, in particular over database data. Along with Andrew Trotman she ran the Data Centric Track at INEX 2010 using an April 10, 2010 dump of the IMDb converted into XML. The specific research question was: Can whole document retrieval approaches outperform focused retrieval on highly structured document collection? This was very much a return to the roots of INEX: Information Retrieval search over a Database in XML.

4.2 Data

There are two kinds of objects in the IMDb collection, movies and persons involved in movies, e.g. actors/actresses, directors, producers and so on. Each object is richly structured. For example, each movie has title, rating, directors, actors, plot, keywords, genres, release dates, trivia, etc.; and each person has name, birth date, biography, filmography, etc. Information about one movie or person is published in one XML file, thus each generated XML file represents a single object, i.e. a movie or person. In total, 4,418,102 XML files were generated, including 1,594,513 movies, 1,872,492 actors, 129,137 directors who did not act in any movies, 178,117 producers who did not direct or act in any movies, and 643,843 other people involved in movies who did not produce or direct nor act in any movies.

4.3 Task and Judgments

Each participating group was asked to create a set of candidate topics, representative of a range of real user needs. Both Content Only (CO) and Content And Structure (CAS)

variants of the information need were requested. In total 30 topics were submitted by 4 institutes of which 28 were used in evaluation.

Participants were permitted to submit up to 10 runs. Each run was permitted to contain a maximum of 1000 results per topic, ordered by decreasing value of relevance. In total 36 runs were submitted by 8 institutes. Only 29 runs were assessed since other runs were submitted after the deadline. Of note is that more runs were submitted than topics, and more institutes submitted runs than that submitted topics. This suggests an increase in interest in the track throughout the year.

Assessment followed standard INEX practice consisting of pooling, participant assessment using the INEX tool, and evaluation using the INEX evaluation tool. In total 26 of the 28 topics were assessed. Runs were scored using MAP, MAiP, and MAgP T2I(300).

4.4 Results

The best run was submitted by Peking University. It used the description and narrative parts of the topics in ranking to identify top documents, then snippet identification techniques to identify good elements in those documents. Of the remaining runs, most did not use structure in ranking (they used the topic title, not the castitle). It is too early to draw conclusions from only one year of experimentation. However, from these runs the result at this stage is that the whole document BM25 retrieval is effective and hard to beat.

4.5 Outlook

The track will be run in 2011. There we expect to see runs that use structure from the query and structure from the documents in order to improve performance. In addition to the ad hoc search task over the richly structured XML data, a new task, faceted search task, is added to examine how to employ faceted search technology to satisfy users complex information needs.

5 Interactive Track

In this section, we will briefly discuss the aims, data, tasks, and results of the INEX 2010 Interactive Track. Further details are in [9].

5.1 Aims

The purpose of the INEX interactive track (iTrack) has been to study searchers interaction with XML-based information retrieval systems, focusing on how end users react to and exploit the potential of systems which provide access to parts of documents in addition to the full documents.

The track was run for the first time in 2004, repeated in 2005, 2006/2007, 2009 and again in 2010. 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.

In this way the participating institutions have gained access to a rich and comparable set of data on user background and user behavior, with a relatively small investment in time and effort. The data collected has been subjected to both qualitative and quantitative analysis, resulting in a number of papers and conference presentations over the years.

5.2 Test Collection

The document collection used for the 2010 iTrack was a slightly modified version of the one used in 2009. A crawl of some 1.5 million records from the book database of the online bookseller Amazon.com has been consolidated with corresponding bibliographic records from the cooperative book cataloguing tool LibraryThing. The records present book descriptions on a number of levels: formalized author, title and publisher data; subject descriptions and user tags; book cover images; full text reviews and content descriptions provided by both publishers and readers.

The experiments were conducted on a Java-based retrieval system built within the ezDL framework (see `http://www.is.inf.uni-due.de/projects/ezdl/` and `http://ezdl.de`), which resides on a server at and is maintained by the University of Duisburg-Essen. The collection was indexed with Apache Solr 1.4, which is based on Apache Lucene. Lucene applies a variation of the vector space retrieval model. Two versions (A and B) were developed for the experiments. The B version of the search system did not allow the user to search in reviews or abstracts, i.e. no query fields for abstract and reviews were available to the user.

5.3 Tasks

For the 2010 iTrack the experiment was designed with two categories of tasks constructed by the track organizers, from each of which the searchers were instructed to select one of three alternative search topics. In addition the searchers were invited to perform one semi-self-generated task. The two task categories were presented in contexts that intended to reflect two different stages of a work task process, where the first set of tasks invited searchers use a broad selection of metadata, and intended to inspire users to create polyrepresentative search strategies, i.e., to use explorative search strategies, which would provide data on query development, metadata type preference and navigation patterns. The second task set intended to simulate searchers in a rather mechanistic data gathering mode. The tasks also intended to represent tasks designed to focus on non-topical characteristics of the books, where information would typically be found in publisher's texts and possible topic-descriptive tags. The self-selected task was intended to function as a control task, where the performance could be compared to the two others.

The experiment was designed to let searchers assess the relevance of books, and they could also simulate the purchase of a book by adding it to a basket. An example of a task in the explorative category 1 would be "*Find controversial books discussing the climate change and whether it is man-made or not.*" An example of a data collection task (Category 2) would be "*Find biographies on athletes active in the 1990's.*" In both cases the tasks were presented in a work task context. At the beginning of the experiment, the participants were asked to fill out a pre-experiment questionnaire. Each task was preceded by a pre-task and concluded by a post-task questionnaire. After all three working tasks had been worked on,

a post-experiment questionnaire was answered. Actions by the system and the participants were recorded by the system and stored in a database.

5.4 Results

Three research groups participated in this year's track: Oslo University College, University of Duisburg-Essen, and University of Glasgow. Data from a total of 147 sessions performed by 49 test subjects were collected from October 2010 to January 2011. Some participants received a EUR 12 Amazon voucher as compensation. Since the data collection phase lasted longer than initially envisioned, the data are currently being analyzed. A primary focus of the analysis will be searchers' choice of sources of information for the completion of the tasks. Amongst the issues which will be looked at is the effect of searchers' topic knowledge and the influence of task types on search procedures and relevance judgements.

6 Question Answering Track

In this section, we will briefly discuss the INEX 2009–2010 QA Track, which aimed to compare the performance of QA, XML or passage retrieval, and automatic summarization systems. Further details are in [10].

6.1 Tasks and Test Collection

We used the Ad Hoc Track's XML enriched Wikipedia based on a late 2008 dump of Wikipedia [11]. A set of 345 questions has been made available:

- Factual questions, related to 2009 ad-hoc topics (151 questions) or extracted from Over-Blog logs (44 questions from `http://en.over-blog.com/`). These questions required a single precise answer to be found in the corpus if it exists.
- Complex questions, also from 2009 topics (85 questions) and Over-Blog (70 questions). Complex questions required a multi-document aggregation of passages with a maximum of 500 words exclusively.

A run is available for factual questions, but this year evaluation has been carried out on the 70 complex Over-Blog questions only. These questions have been selected such that there exists at least a partial answer in the Wikipedia 2008. We have mixed these questions with others from Yahoo! Answers website (`http://answers.yahoo.com/`).

We considered three different types of questions: short_single, short_multiple and long. Those labeled short_single or short_multiple are 195 and both require short answers, *i.e.* passages of a maximum of 50 words together with an offset indicating the position of the answer. Short_single questions should have a single correct answer, whereas multiple type questions will admit multiple answers. For both short types, participants had to give their results as a ranked list of maximum 10 passages from the corpus together with an offset indicating the position of the answer. Long type questions require long answers up to 500 words that must be self-contained summaries made of passages extracted from the INEX 2009 corpus.

A state of the art IR engine powered by Indri was made available to participants. It allowed the participation of seven summarization systems for the first time at INEX. These

systems only considered long type answers and have been evaluated on the 2010 subset. Only two standard QA systems participated to the factual question sub-track.

6.2 Results

Only long answer evaluation is presented here. Long answers have been evaluated based on Kullback Leibler (KL) divergence between n-gram distributions. The informative content of the long type answers is evaluated by comparing the several n-gram distributions in participant extracts and in a set of relevant passages selected manually by organizers.

All seven participants for long type questions used the provided IR engine and generated summaries by sentence extraction. This helps readability even if it does not ensure general coherence. The standard deviation among systems KL divergences varies. The ten questions minimizing standard deviation and, therefore, getting most similar answers among systems are those containing at least one named entity that refers to a Wikipedia page. Therefore, the systems mostly built their answer based on the textual content of this page and KL divergence is not accurate enough to discriminate among them. On the contrary, the questions that maximize standard deviation and have the greatest impact in the ranking of the systems are non-encyclopedic ones and do not refer to particular Wikipedia pages. Meanwhile partial answers exist in the Wikipedia but they are spread among several articles.

6.3 Conclusion and Outlook

We will keep on addressing real-world focused information needs formulated as natural language questions using special XML annotated dumps of the Wikipedia, but we will mix questions requiring long answers and those not. We shall build and make available a new dump of the Wikipedia adapted for QA where only document structure, sentences and named entities will be annotated. Moreover, the information concerning the short/long type of expected answers will be removed from the input question format. Given a question, the decision whether the answer should be rather short or long is then left to the system. For each question, participants will have to guess the context category (for example, "expert," "pupil," "journalist," ...), provide a context as a query-oriented readable summary and a list of possible answers. Real questions will be selected from OverBlog website logs and Twitter.

7 Link the Wiki Track

In this section, we will briefly discuss the collection, the task, and the submissions and results of the INEX 2010 Link the Wiki Track. Further details are in [13].

7.1 Collection

In 2010 the Link the Wiki track moved away from the Wikipedia and in its place used the Te Ara collection supplied by the New Zealand Ministry of Culture and Heritage. Te Ara is an online encyclopaedia without hypertext links and as such forms an ideal platform for link discovery. The collection consists of 36,715 documents totalling 48MB in size.

Linking Te Ara is more complex than linking the Wikipedia for many reasons. The articles are often digital narratives and in being so do not represent entities—the controlled

vocabulary of the Wikipedia is not present. A standard title-matching algorithm was thus not expected to be effective. The collection also does not contain any links and so link-mining algorithms such as that of Kelly and Clarke [6] could not be used.

7.2 Task and Results

The task was to link the entire encyclopaedia to itself. Runs consisted of up-to 50 anchors per document with up-to 5 anchors per link. Only two institutes, University of Otago and Queensland University of Technology, submitted runs, with Otago submitting 24 and QUT submitting 5.

A set of documents was randomly chosen for manual assessment. The collection was ordered on the number of links in each document. This was then divided into 10 deciles. A work-set was constructed by randomly selecting one document from each decile. The links for these documents were then pooled. Seven (non-overlapping) work-set pools were assessed to completion resulting in a total of 70 assessed documents. For 18 of the assessed documents had no relevant links were found. The mean number of relevant targets per topic was 8.8 and the mean number of non-relevant targets per topic was 274.6. Topic 2919 had the most relevant targets (97).

As is the informal convention at INEX, the metrics for the Link-the-Wiki track in 2010 were not published before the runs were submitted. As is also the informal convention, the metric changed in 2010. In a run it is possible (and correct) to identify more than one anchor targeting the same document. It is also possible and correct to identify more than one target per anchor. Consequently metrics based on recall (such as Mean Average Precision, MAP) are meaningless. If there is only one relevant target document, but the link-discovery algorithm identifies two different anchors for that target then what is the recall? This did happen in the submitted runs. The runs were de-duplicated by taking only the highest ranking instance of a target for a given topic and ignoring other instances of the target. The relevance of each target was then determined using the manual assessments. Then the score for each position in the results list was assigned a 1 if any target for that anchor was relevant and 0 otherwise. Mean Average Precision was then computed from this. This approach gives a highly optimistic evaluation because the run has 5 trials at each point in the results list.

The best run was submitted by Otago. That run constructed anchors as maximal sequences of consecutive words delineated by stop words or punctuation. The source document is then used to identify relevant target documents, and the anchor used to identify relevant headings in those documents.

7.3 Outlook

In 2011 the track organisers will be examining cross-lingual link discovery in the Wikipedia. This new experiment is the basis of the CrossLink track at NTCIR 9.

8 Relevance Feedback Track

In this section, we will briefly discuss the INEX 2010 Relevance Feedback Track. Further details are in [3].

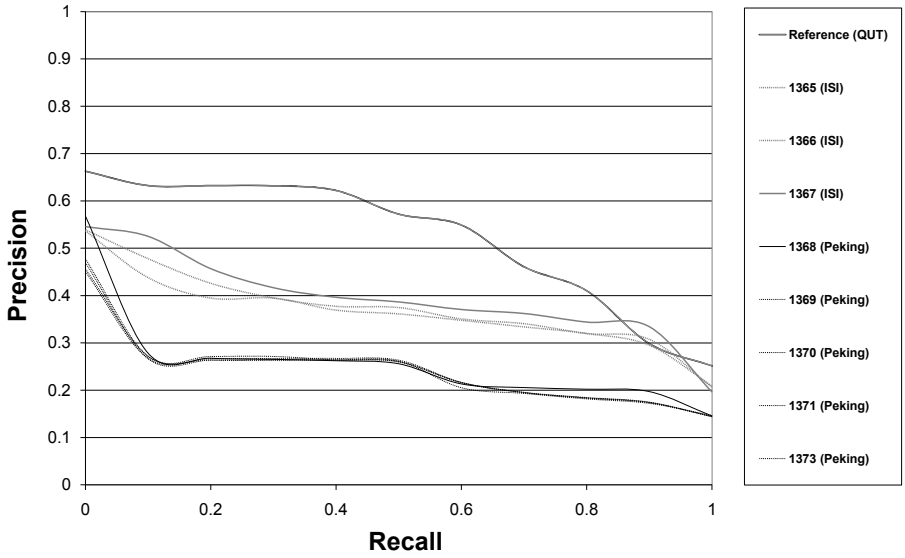

Figure 1: Recall-precision comparison of Relevance Feedback Modules

8.1 Overview

The Relevance Feedback track investigates the simulation of a user interacting with an information retrieval system, searching for a number of different topics. The user provides feedback to the system in the form of nominating relevant passages in the retrieved texts, which is then used by the system to provide improved results to the user. This is an alternative approach to traditional methods of relevance feedback using document level feedback. The quality of the results this user receives is then used to evaluate the approach.

8.2 Track Design

Participating organisations submit a *relevance feedback module* (RFM) in the form of a Java library. Each RFM implements a particular algorithm for ranking and reranking a list of provided documents according to feedback passed to it. The RFMs are each evaluated by an *evaluation platform* (EP) that links the RFMs, provides them with a hitherto unknown set of documents and simulates a user making use of the system, looking through the documents returned by the RFM and providing relevance judgments based on preexisting relevance assessments. The RFMs are then scored based on how well they returned relevant information to the simulated user.

8.3 Submissions

Including the reference RFM, available in source and binary form before the track was run, there were a total of nine submissions from three participating organisations. QUT submitted the aforementioned reference module, which applied a simplistic implementation of Rocchio and was tuned to an earlier data set provided along with the module. The Indian Statistical Institute (ISI) submitted a relevance feedback algorithm that was designed around finding non-overlapping word windows in the relevant passages and modifying the query to include

these terms. Peking University submitted an algorithm that used a Rocchio-based algorithm revised to include negative feedback and criterion weight adjustments.

8.4 Results

The results, which were evaluated using trec_eval, are shown in Figure 1. The relatively flat precision over recall curve shows that the incremental feedback is clearly effective and improving over the ranking. Obviously, the results with the feedback applied for each of the modules is superior to the performance of the modules in ranking the documents before any feedback is given. We refer to [3] for further detail and comparisons.

9 Web Service Discovery Track

In this section, we briefly discuss the INEX 2010 Web Service Discovery Track. Further details are in [12].

9.1 WSDL Collection and Topics

The track, introduced in 2010, addresses discovery of Web services based on descriptions provided in Web Services Description Language (WSDL). The track used a collection of 1,738 WSDL documents that were directly crawled from real-world public Web services indexed by the Google search engine (only valid WSDL1.1-compliant descriptions were retained).

The five active participating groups contributed topics which aimed to be representative of a range of realistic web service discovery needs. The topics used the same format as the ad hoc track, including a title (or keyword query), a structured query, topic description, and narrative explaining how relevance will be judged for the topic. For example, the topic *airline flights* represents a query one might pose if searching for a web service that can find flight details given an airline flight number. Of the topics submitted, 25 were used in submissions for the track in 2010; however participants only ended up judging 20 of these topics.

9.2 Results

Submission were allowed in the same formats as the ad hoc track: XML elements using XPath syntax, or passages using file offsets and lengths, or ranges of elements. A total of 15 runs were received from the five participating groups, however these runs were only document level runs. Some minor formatting corrections were done on submissions, and three runs were excluded from the final evaluation due to more serious errors (such as duplicate answers).

All 15 runs were used to contribute documents into the assessment pool. The assessment pool contained approximately 100 documents for each topic. A modified version of evaluation tool for the ad hoc track was used, which displayed the XML markup of the WSDL documents along the content of the elements. Assessors needed to see the XML structure to judge whether a document (or parts of a document) were relevant, since in many cases the WSDL document contained little or no text content in the XML elements.

Since only document level runs were submitted, evaluation was only performed for document retrieval using mean average precision. The best performing run, which was one of runs submitted from Kasetsart University, had a mean average precision of 0.3469.

9.3 Outlook

Following the development of an initial test collection in 2010 which allows comparative retrieval experiments, we hope to get more groups participating in 2011. A goal of future years in this track, will be to allow topics to include a description of workflow, in order to find web services that can meet steps within a larger process.

10 XML Mining Track

In this section, we will briefly discuss the INEX 2010 XML Mining track. Further details are in [4].

10.1 Aims and Tasks

The aims of the INEX 2010 XML Mining track are: (1) studying and assessing the potential of data mining (DM) techniques for dealing with generic DM tasks in the structured domain i.e. classification and clustering of XML documents; and (2) evaluating clustering approaches in the context of XML information retrieval. The INEX 2010 XML Mining track included two tasks: (1) unsupervised clustering task and (2) semi-supervised classification task.

The clustering task in INEX 2010 continued to explicitly test the cluster hypothesis, which states that documents that cluster together have a similar relevance to a given query. It uses manual query assessments from the INEX 2010 Ad Hoc track. If the cluster hypothesis holds true, and if suitable clustering can be achieved, then a clustering solution will minimize the number of clusters that need to be searched to satisfy any given query. More precisely, this is a multiple label clustering task where the goal was to associate each document to a single or multiple clusters in order to determine the quality of cluster relative to the optimal collection selection goal, given a set of queries.

The classification task in INEX 2010 focused on evaluating the use of external structure of the collection i.e the links between documents along with the content information and the internal structure of XML documents for classifying documents into multiple categories. More precisely, this is a a multiple label classification task where the goal was to find the single or multiple categories of each document. This task considers a transductive context where, during the training phase, the whole graph of documents is known but the labels of only a part of them are given to the participants.

10.2 Test Collection

The XML Mining task used a subset of 2.7 million English Wikipedia XML documents, their labels, a set of information needs (i.e., the Ad Hoc track queries), and the answers to those information needs (i.e., manual assessments from the Ad Hoc track). A 146,225 document subset was used as a data set for the clustering and classification tasks. The subset is determined by the reference run used for the ad hoc track. Using the reference run reduced the collection from 2,666,190 to 146,225 documents.

The clustering evaluation uses Ad Hoc relevance judgments for evaluation and most of the relevant documents are contained in the subset. The reference run contains approximately 90 percent of the relevant documents.

A new approach to extracting document category labels was taken this year. The Wikipedia category graph was processed to extract 36 categories for documents. This is completed by searching for shortest paths through the graph.

In order to enable participation with minimal overheads in data-preparation the collection was pre-processed to provide various representations of the documents such as a bag-of-words representation of terms and frequent phrases in a document, frequencies of various XML structures in the form of trees, links, named entities, etc.

10.3 Evaluation and Results

Participants were asked to submit multiple clustering solutions containing different numbers of clusters such as 50, 100, 200, 500 and 1,000. For the clustering task, the participants had submitted a cluster index(es) for each document of the collection set. The clustering solutions were evaluated by two means. Firstly, we utilize the classes-to-clusters evaluation which assumes that the classification of the documents in a sample is known (i.e., each document has a class label). For each submitted cluster, we have computed the standard Purity, Entropy and Normalized Mutual Information (NMI) scores, indicating how well a clustering solution matches the ground truth. Secondly, we evaluate clustering solutions against the optimal collection selection goal. A total of 69 topics used in Ad Hoc track were utilized to evaluate the quality of clusters generated on the 144,265 XML Mining document subset. The Normalized Cluster Cumulative Gain (NCCG) is used to calculate the score of the best possible collection selection according to a given clustering solution.

A total of three research teams have participated in the INEX 2010 clustering task. As expected, as the cluster numbers dividing the data set increases, performance of all the teams based on the Purity, Entropy and NMI scores increase and based on the NCCG score decreases. Analysis of results by various submissions show that the most recall comes from the first few clusters confirming the hypothesis that a good clustering solution tends to (on average) group together relevant results for ad-hoc queries.

For classification, we have asked the participants to submit multiple category for each of the documents of the testing set. We have then evaluated how much the categories found by the participants correspond to the real categories of the documents. For each category, we have computed a F1 score that measures the ability of a system to find the relevant categories. Two different teams had participated to the task. All teams except were able to achieve micro and macro F1 between 45-55%. Detailed results are given in [4].

11 Envoi

This complete our walk-through of the seven tracks of INEX 2010. The tracks cover various aspects of focused retrieval in a wide range of information retrieval tasks. This report has only touched upon the various approaches applied to these tasks, and their effectiveness. The formal proceedings of INEX 2010 are being published in the Springer LNCS series [5]. This volume contains both the track overview papers, as well as the papers of the participating groups. The main result of INEX 2010, however, is a great number of test collections that can be used for future experiments.

The INEX 2010 Workshop was held in the Netherlands, on 13–15 December 2010, with lively reports on the INEX tracks and extensive discussion on what to do at INEX 2011.

The Dutch Association for Information Science (*Werkgemeenschap Informatiewetenschap*) sponsored a best student award, which was presented to Ning Gao (Master Student at Peking University) for her paper entitled "Combining Strategy for XML Retrieval."

INEX 2011 will see some exciting changes. Test collections for some tracks, including Ad Hoc search against Wikipedia, seem completed and are available as solid benchmarks for use and reuse. INEX 2011 will proudly continue pushing research boundaries, with a wide range of new tasks including Social Search, Faceted Search, and Snippet Retrieval.

References

[1] P. Arvola, S. Geva, J. Kamps, R. Schenkel, A. Trotman, and J. Vainio. Overview of the INEX 2010 ad hoc track. In Geva et al. [5].

[2] P. Arvola, J. Kekäläinen, and M. Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13:460–484, 2010.

[3] T. Chappell and S. Geva. Overview of the INEX 2010 relevance feedback track. In Geva et al. [5].

[4] C. M. De Vries, R. Nayak, S. Kutty, S. Geva, and A. Tagarelli. Overview of the INEX 2010 XML mining track: Clustering and classification of XML documents. In Geva et al. [5].

[5] S. Geva, J. Kamps, R. Schenkel, and A. Trotman, editors. *Comparative Evaluation of Focused Retrieval : 9th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2010)*, LNCS. Springer, 2011.

[6] K. Y. Itakura and C. L. A. Clarke. University of waterloo at INEX2007: Adhoc and link-the-wiki tracks. In *Focused Access to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, pages 417–425, 2008.

[7] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. INEX 2007 evaluation measures. In *Focused Access to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, pages 24–33, 2008.

[8] G. Kazai, M. Koolen, J. Kamps, A. Doucet, and M. Landoni. Overview of the INEX 2010 book track: Scaling up the evaluation using crowdsourcing. In Geva et al. [5].

[9] N. Pharo, T. Beckers, R. Nordlie, and N. Fuhr. Overview of the INEX 2010 interactive track. In Geva et al. [5].

[10] E. SanJuan, P. Bellot, V. Moriceau, and X. Tannier. Overview of the INEX 2010 question answering track. In Geva et al. [5].

[11] R. Schenkel, F. M. Suchanek, and G. Kasneci. YAWN: A semantically annotated Wikipedia XML corpus. In *12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2007)*, pages 277–291, 2007.

[12] J. A. Thom and C. Wu. Overview of the INEX 2010 web service discovery track. In Geva et al. [5].

[13] A. Trotman, D. Alexander, and S. Geva. Overview of the INEX 2010 link the wiki track. In Geva et al. [5].

[14] A. Trotman and Q. Wang. Overview of the INEX 2010 data centric track. In Geva et al. [5].