Report on INEX 2008

Gianluca Demartini **Ludovic Denoyer** **Antoine Doucet** **Khairun Nisa Fachry**
Patrick Gallinari **Shlomo Geva** **Wei-Che Huang** **Tereza Iofciu**
Jaap Kamps **Gabriella Kazai** **Marijn Koolen** **Monica Landoni**
Ragnar Nordlie **Nils Pharo** **Ralf Schenkel** **Martin Theobald**
Andrew Trotman **Arjen P. de Vries** **Alan Woodley** **Jianhan Zhu**

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 2008 evaluation campaign, which consisted of a wide range of tracks: Ad hoc, Book, Efficiency, Entity Ranking, Interactive, QA, Link the Wiki, and XML Mining.

1 Introduction

Traditional search engines identify whole documents that are relevant to a user's information need, the task of locating the relevant information within the document is left to the user. Next generation search engines will perform both tasks: they will identify relevant parts of relevant documents. A search engine that performs such a task is referred to as focused and the discipline is known as Focused Retrieval. The main goal of INEX is to promote the evaluation of focused retrieval 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 2008 evaluation campaign consisted of a wide range of tracks:

Ad hoc Track Investigating the effectiveness of XML-IR and Passage Retrieval for three ad hoc retrieval tasks (Focused, Relevant in Context, Best in Context).

Book Track Investigating techniques to support users in reading, searching, and navigating full texts of digitized books.

Efficiency Track Investigating the trade-off between effectiveness and efficiency of ranked XML retrieval approaches on real data and real queries.

Entity Ranking Track Investigating entity retrieval rather than text retrieval: 1) Entity Ranking, 2) Entity List Completion.

Interactive Track (iTrack) Investigating the behavior of users when interacting with XML documents, and retrieval approaches which are effective in user-based environments.

Link-the-Wiki Track Investigating link discovery between Wikipedia documents, both at the file level and at the element level.

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

In addition, there were initial steps to launch a *Question Answering* track, investigating how technology for accessing semi-structured data can be used to address interrogative information needs, and a *Wikipedia Vandalism* track, trying to predict edit reversal in Wikipedia.

In the rest of this paper, we discuss the aims and results of the INEX 2008 tracks in relatively self-contained sections: the Ad Hoc track (Section 2), the Book track (Section 3), the Efficiency track (Section 4), the Entity Ranking track (Section 5), the Interactive track (Section 6), the Link the Wiki track (Section 7), and the XML Mining track (Section 8).

2 Ad Hoc Track

In this section, we will briefly discuss the aims of the Ad Hoc track, its tasks and setup, the used measures and results, and try to formulate clear findings. Further details are in [14].

2.1 Aims and Tasks

The Ad Hoc Track at INEX studies the adhoc retrieval of XML elements or passages. In information retrieval (IR) literature, adhoc retrieval is described as a simulation of how a library might be used, and it involves the searching of a static set of documents using a new set of topics. While the principle is the same, the difference for INEX is that the library consists of XML documents, the queries may contain both content and structural conditions and, in response to a query, arbitrary XML elements may be retrieved from the library.

The general aim of an IR system is to find *relevant information* for a given topic of request. In the case of XML retrieval there is, for each article containing relevant information, a choice from a whole hierarchy of different elements or passages to return. Hence, within XML-IR, we regard as *relevant results* those results that both

- contain relevant information (the result exhaustively discusses the topic), but
- contain as little non-relevant information as possible (the result is specific for the topic).

In traditional document retrieval only the first condition is applied. The INEX 2008 measures are solely based on the retrieval of highlighted text. We simplify all INEX tasks to highlighted text retrieval and assume that systems should return all, and only, highlighted text. We then compare the characters of text retrieved by a search engine to the number and location of characters of text identified as relevant by the assessor. For best in context (discussed below) we use the distance between the best entry point in the run and that identified by an assessor.

The INEX 2008 Ad Hoc Track featured three tasks: For the *Focused Task* a ranked-list of non-overlapping results (elements or passages) must be returned. It is evaluated at early precision relative to the highlighted (or believed relevant) text retrieved. For the *Relevant in Context Task* non-overlapping results (elements or passages) must be returned, these are grouped by document. It is evaluated by mean average generalized precision where the generalized score per article is based on the retrieved highlighted text. For the *Best in Context Task* a single starting point (element's starting tag or passage offset) per article must be returned. It is also evaluated by mean average generalized precision but with the generalized score (per article) based on the distance to the assessor's best-entry point.

2.2 Test Collection

INEX 2008 used the Wikipedia XML Corpus based on the English Wikipedia in early 2006, containing a total of 659,338 Wikipedia articles [3]. On average an article contains 161 XML nodes. 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.

INEX has been pioneering peer-topic creation and peer-assessments since 2002. At INEX 2008, a total of 135 ad hoc search topics were created by participants. In addition, 150 queries were derived from a proxy-log. A total of 86 topics has a structured CAS query, for the other topics a default CAS query was added.

The topics were assessed by participants following precise instructions. The assessors used the new GPXrai assessment system that assists assessors in highlight relevant text. Topic assessors were asked to mark all, and only, relevant text in a pool of documents. After assessing an article with relevance, a separate best entry point decision was made by the assessor. The relevance judgments were frozen on October 22, 2008. At this time 70 topics had been fully assessed. Moreover, 11 topics were judged by two separate assessors, each without the knowledge of the other. All official results refer to the 70 topics with the judgments of the first assigned assessor, which is typically the topic's original author.

The main INEX 2008 test-collection consists of the 70 human created and judged topics, and the specific measures to evaluate the three tasks. In addition, trec-style qrels have been derived—treating every article that contains highlighted text as relevant—for evaluating document retrieval effectiveness on the Wikipedia. This results in an attractive document retrieval test collection using freely available documents in a non-news genre. Moreover, trec-style qrels are also available for 125 topics derived from the proxy-log—treating every clicked article as relevant. These topics shed light on the similarities and differences between traditional IR test collections and the data collected in log files.

2.3 Results

There were a total of 163 official submissions by 27 groups, distributed evenly across the three tasks. We report here on the main observations and findings, and refer for a detailed discussion of the results and the top scoring runs to [14].

When examining the relative effectiveness of CO and CAS we found that for all tasks the best scoring runs used the CO query. This is in contrast with earlier results showing that structural hints can help promote initial precision. Part of the explanation may be in the low number of CAS submissions (28) in comparison with the number of CO submissions (108). Only 39 of the 70 judged topics had a non-trivial CAS query, and the majority of those CAS queries made only reference to particular tags and not on their structural relations. This may have diminished the value of the CAS query in comparison with earlier years.

Given the efforts put into the fair comparison of element and passage retrieval approaches, the number of passage submissions was disappointing. Eighteen submissions used ranges of elements or FOL passage results, whereas 118 submissions used element results. Consistent with earlier results on using passage-level evidence for XML element retrieval, we saw that the passage based approaches were competitive, but not superior to element based approaches.

As in earlier years, we saw that article retrieval is reasonably effective at XML-IR. For all the tasks there were article-only runs that ranked relatively high. When looking at the

article rankings inherent in all Ad Hoc Track submissions, i.e., evaluate them as traditional document retrieval, we saw that best article rankings were obtained from runs with element or passage results. This suggests that element-level or passage-level evidence is still valuable for article retrieval. When comparing the system rankings in terms of article retrieval with the system rankings in terms of the INEX retrieval tasks, over the exact same topic set, we see a reasonable correlation especially for the two "in context" tasks. The systems with the best performance for the ad hoc tasks, also tend to have the best article rankings.

Since finding the relevant articles can be considered a prerequisite for XML-IR, this should not come as a surprise. In addition, the Wikipedia's encyclopedic structure with relatively short articles covering a single topic results in relevant articles containing large fractions of relevant text (with a mean of 55% of text being highlighted). While it is straightforward to define tasks and measures that strongly favor precision over recall, a more natural route would be to try to elicit more focused information needs that have natural answers in short excerpts of text.

When we look at a different topic set derived from a proxy log, and a shallow set of clicked pages rather than a full-blown IR test collection, we see notable differences. Given the low number of relevant articles (1.8 on average) compared to the ad hoc judgments (70 on average), the clicked pages focus exclusively on precision aspects. This leads to a different system ranking, although there is still some agreement on the best groups. The differences between these two sets of topics require further analysis.

2.4 Outlook

Finally, the Ad Hoc Track had two main research questions. The first main research question was the comparative analysis of element and passage retrieval approaches, hoping to shed light on the value of the document structure as provided by the XML mark-up. Although the number of non-element retrieval runs submitted is too low to draw any definite conclusions, we found that the best performing system used predominantly element results, providing evidence for the usefulness of the document structure. The second main research question was to compare focused retrieval directly to traditional article retrieval. We found that the best scoring Ad Hoc Track submissions also tend to have the best article ranking, but that the best article rankings were generated using element-level evidence.

Building on the success of the Ad Hoc track at INEX 2008, there will be a number of exciting changes at INEX 2009. First and foremost, there will be a new collection. Based on a 2009 dump of the English Wikipedia, with over 2.5 million articles and billions of elements. This will present a significant test for scaling the INEX infrastructure as well as the systems of individual participants. Second, there will be additional efforts during topic creation that aim to promote more focused information requests. For example, the collection will enriched with semantic annotation that will allow information needs to be naturally cast as structured queries. Third, although ensuring comparability over years suggests running the same tasks on the new collection, there is active debate on some variant tasks that highlight other aspects of XML-IR.

3 Book Track

In this section, we briefly discuss the Book track. For further details, we refer to [17].

3.1 Goals and Setup

Now in its second year, the Book Track [17] focused on three themes of interest relevant to information retrieval (IR), human computer interaction (HCI), digital libraries (DL), and eBooks: a) IR techniques for searching collections of digitized books, b) users' interactions with eBooks, and c) mechanisms to increase accessibility to the content of digitized books. Based on these, four tasks were defined and investigated: 1) The *Book Retrieval* (BR) task aimed to compare traditional document retrieval methods with domain-specific techniques exploiting book-specific features, such as the back of book index or associated metadata like library catalogue information, framed within the user task of building a reading list for a given topic, 2) the *Page in Context* (PiC) task aimed to test the value of applying focused retrieval approaches to books where users expect to be pointed directly to relevant book parts, 3) the *Structure Extraction* (SE) task aimed to evaluate automatic techniques for deriving structure from layout and OCR for building hyperlinked table of contents (ToCs) for digitized books, and 4) the *Active Reading* task (ART) aimed to explore suitable user interfaces enabling annotation, review and summary across multiple books.

3.2 Test Collection

A total of 54 organisations registered for the track, of which 15 took part actively throughout the year, contributing topics, runs, or relevance judgements to the test collection.

The test collection is based on 50,239 digitized out-of-copyright books (totaling 400GB), provided by Microsoft Live Search and the Internet Archive. These include history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry. The full text of the books is marked up in an XML format referred to as BookML, developed by the Document Layout Team of Microsoft Development Center Serbia, which contains, e.g., markup for table of contents entries. 50,099 of the books also comes with an associated MAchine-Readable Cataloging (MARC) record that contains publication (author, title, etc.) and classification information. In addition to the full corpus, a reduced version (50GB, or 13GB compressed) was also made available, where word markups (incl. word coordinates) were removed. Both the BR and PiC tasks built on the full corpus, while in ART participants could select up to 100 books to use in their user studies, and the SE task used a different set of 100 books for which JPEG page images and the original OCR files (in DjVu XML, essentially with only page level structure) were distributed to participants.

In 2008, 40 new content-only (CO) topics (ID: 31-70) were contributed by participants, which were merged with the 30 CO topics created last year for the PiC task (ID: 1-30). The combined set was then used both for the BR and PiC tasks.

Relevance assessments were collected using an online Book Search System, available at `http://www.booksearch.org.uk`, developed by Microsoft Research Cambridge, which allowed participants to search, browse, read and annotate books in the test collection. Assessments were gathered through a game called the Book Explorers' Competition, which was modeled based on two competing roles: explorers vs reviewers. An explorer's task was to locate and mark relevant content. Reviewers then checked the quality of the explorers' work by providing their own assessments. In addition, both explorers and reviewers judged the relevance of books on a six-point scale. The collection of relevance assessments was frozen on 25 February 2009. In total, 3,674 unique books and 33,120 unique pages were judged across

29 topics, and 1,019 highlight boxes were drawn by 17 assessors. For more details on the collected data, please refer to [17].

3.3 Results

3.3.1 Book Retrieval and Page in Context Tasks

For the evaluation of the BR and PiC tasks, we used trec_eval v8.1 and separate book-level and page-level relevance assessment sets (qrels), where multiple relevance labels assigned by multiple assessors were averaged. The ranking of books in both BR and PiC tasks was evaluated as traditional document retrieval. The ranking of book parts in the PiC task was evaluated at page level for each book, treating each page as a document, and then averaging over the run. We summarise below the main findings, but note that since the qrels vary greatly across topics, these should be treated more as preliminary observations.

For the BR task, 18 runs were submitted by 4 groups. Participants experimented with various techniques, e.g., using book content vs. MARC record information, ranking books by document score vs. best element score, or ranking books by the percentage of pages retrieved, as well as incorporating Wikipedia evidence. The best performing run (by MAP) was a run submitted by RMIT, which ranked books by the percentage of pages retrieved using BM25 over a page level index (MAP=0.1056). The general conclusion, however, for the other 3 groups' experiments was that the simple book content based baseline performed better than any attempts to combine book-specific evidence to improve performance. This suggests that there is still plenty to be done in discovering suitable ranking strategies for books.

For the PiC task, 13 runs were submitted by 2 groups. Participants mostly experimented with ways of combining document and element level scoring methods. The best performing run was submitted by the University of Amsterdam, who found that while focused methods were able to locate relevant text within books, page level evidence was of limited use without the wider context of the whole book.

3.3.2 Structure Extraction Task

For the evaluation of the SE task, the ToCs generated by participants were compared to a manually built ground-truth, created by hired assessors, using a structure labeling tool built by Microsoft Development Center Serbia. Precision was defined as the ratio of the total number of correctly recognized ToC entries and the total number of returned ToC entries; and recall as the ratio of the total number of correctly recognized ToC entries and the total number of ToC entries in the ground-truth.

7 runs were submitted by 2 groups, implementing two very different approaches. The best performance (by the F-measure, the harmonic mean of precision and recall), was obtained by the Microsoft Development Center Serbia team ($F = 53.47\%$), who extracted ToCs by first recognizing the page(s) of a book that contained the printed ToCs. The other group relied on title detection within the body of a book and achieved a score of $F = 10.27\%$.

3.3.3 Active Reading Task

The main aim of ART is to explore how hardware or software tools for reading eBooks can provide support to users engaged with a variety of reading related activities, such as fact

finding, memory tasks or learning. The goal of the investigation is to derive user requirements and consequently design recommendations for more usable tools to support active reading practices for eBooks. This is done by running a comparable but individualized set of studies, all contributing to elicit user and usability issues related to eBooks and e-reading. Because of its novelty, it took a while to involve and engage researchers in ART. Studies run by 2 participating groups are still ongoing, and thus we do not yet have results to report. We are continuing ART in 2009 and plan to work toward raising awareness and interest in related communities not yet involved in INEX.

3.4 Conclusions and Outlook

The Book Track in 2008 has attracted a lot of interest and has grown to double the number of participants from 2007. However, active participation remained a challenge for most groups due to the high initial setup costs (e.g., building infrastructure). Nonetheless, a lot has been achieved this year. The most significant result is an established infrastructure for the evaluation of the various tasks. These include evaluation mechanisms, measures, user study methodologies, and ground-truth building methods and systems. The latter presented one of the biggest challenges due to the huge effort required. We devised a collective relevance gathering method, which we implemented as an online game. We found this method feasible and reliable [18], but one that requires a larger community to support it, i.e., >> 17 assessors. To address this, we are currently looking at using Amazon's Mechanical Turk service, as well as investigating the possibility of opening up the Book Search System and allowing any users to create their own topics and saving their searches and book annotations for these.

For INEX 2009, we plan to run modified versions of the same tasks. The SE task will run both at INEX 2009 and at ICDAR 2009 (International Conference on Document Analysis and Recognition) with a set of 1,000 books. The BR task will be shaped around the user task of compiling a reading list for selected Wikipedia articles, while we aim to expand the PiC tasks to tree retrieval [1]. ART is continuing into 2009.

4 Efficiency Track

In this section, we discuss the goals, general setup and results of the Efficiency Track that was newly introduced to INEX 2008. For further details, we refer to [21].

4.1 Overview

The new INEX Efficiency Track provides a common forum for the evaluation of both the *effectiveness* and *efficiency* of XML ranked retrieval approaches on *real data* and *real queries*. As opposed to the purely synthetic XMark or XBench benchmark settings that are still prevalent in efficiency-oriented XML retrieval tasks, the Efficiency Track continues the INEX tradition using a rich pool of manually assessed relevance judgments for measuring retrieval effectiveness. Thus, one of the main goals is to attract more groups from the DB community to INEX, being able to study effectiveness/efficiency trade-offs in XML ranked retrieval for a broad audience from both the DB and IR communities. The Efficiency Track significantly extends the Ad-Hoc Track by systematically investigating different types of queries and retrieval scenarios, such as classic ad-hoc search, high-dimensional query expansion settings,

and queries with a deeply nested structure (with all topics being available in both the NEXI-style CO and CAS formulations, as well as in their XPath 2.0 Full-Text counterparts).

Just like the Ad-Hoc Track, the Efficiency Track used the 2007 version of the INEX-Wikipedia collection [3], an XML version of English Wikipedia articles initially introduced for INEX 2006 and slightly revised in 2007. Although this 4.38 GB XML-ified Wikipedia collection is not particularly large from a DB point-of-view, it has a rather irregular structure with many deeply nested paths, which will be particularly challenging for traditional DB-style approaches, e.g., using path summaries. There is no DTD available for INEX-Wikipedia.

4.2 Topic Types

One of the main goals to distinguish the Efficiency Track from traditional Ad-Hoc retrieval was to cover a broader range of query types than the typical NEXI-style CO or CAS queries, which are mostly using either none or only very little structural information and only a few keyword conditions over the target element of the query. Thus, two natural extensions were 1) to extend given Ad-Hoc queries with high-dimensional query expansions, and 2) issue a specific call for new topics to all participants, aiming to increase the amount of structural query conditions without sacrificing IR aspects in processing these queries. In summary, the Efficiency Track focused on the following types of topics, each representing different challenges for efficient and effective retrieval:

Type (A) Topics: 540 topics (no. 289–828) were taken over from previous Ad-hoc Track settings used in 2006–2008, which constituted the major bulk of topics used for the Efficiency Track. These topics represent classic, Ad-Hoc-style, focused passage or element retrieval (similar to the INEX Ad-Hoc Focused task 2006–2008) over a combination of NEXI CO and CAS queries. Topic ids were taken over from the Ad-Hoc track, thus allowing for the reuse of assessments.

Type (B) Topics: 21 topics (no. 829–849) were derived from interactive, feedback-based query expansion runs, kindly provided by the Royal School of Library and Information Science, Denmark, investigated in the context of the INEX Interactive Track 2006. These CO topics were intended to simulate high-dimensional query expansion settings with up to 112 keywords (topic no. 844), which cannot be evaluated in a conjunctive manner and are expected to pose a major challenge to any kind of search engine and evaluation strategy. Respective expansion runs have been submitted by RSLIS also to the 2006 Ad-Hoc track, such that relevance assessments for these topics are available from the INEX 2006 Ad-Hoc track assessments.

Type (C) Topics: 7 new topics (no. 850–856) were newly developed and submitted by Efficiency Track participants. These topics represent high-dimensional, structure-oriented retrieval settings over a DB-style set of CAS queries, with deeply nested structure but only a few keyword conditions. Assessments were originally intended to get accomplished by Efficiency Track participants as well, but were then skipped due to the low amount of newly proposed type (C) topics and the low respective impact on overall result effectiveness as compared to the more than 500 Ad-Hoc topics that already come readily assessed. The evaluation of run-times however remains very interesting over this structure-enhanced set of type (C) topics as well.

The reuse of type (A) and (B) led to 308 topics for which assessments from the INEX 2006–2008 Ad-hoc Tracks are readily available. An additional conversion to the new 2008 version of the `inex_eval` tool and the (passage-based) assessments format was needed to incorporate the 2008 assessment files (QRels) and has been made available online for download from the track homepage at `http://www.inex.otago.ac.nz/tracks/efficiency/efficiency.asp`.

4.3 Tasks and Metrics

The Efficiency Track particularly encouraged the use of top-k style query engines. The result submission format included options for marking runs as top-15, top-150, and top-1,500 (the latter corresponding to the traditional Ad-hoc submission format), using either a *Focused* (i.e., non-overlapping), *Thorough* (incl. overlap), or *Article* retrieval mode. Automatic runs may use either the title field, including the NEXI CO, CAS, or XPATH titles, additional keywords from the narrative or description fields, as well as automatic query expansions if desired. As opposed to the Ad-Hoc Track, reconsidering a Thorough retrieval mode (as used initially in INEX 2003–2005) intentionally allowed for overlapping elements to be returned, since removing overlap may mean a substantial burden for different systems.

To assess the quality of the retrieved results, the Efficiency Track applied the same metrics as used in the Ad-Hoc track. Runs in *Focused* or *Article* mode were evaluated with the interpolated precision metric [15], using the evaluation toolkit from INEX 2008; the assessments for the topics from 2006 and 2007 have been converted to the new Qrel-based format. Runs in *Thorough* mode were evaluated with the precision-recall metric as implemented in `inex_eval` [9] after converting the Qrels from 2008 to the old XML-based assessment format.

4.4 Results and Conclusions

We received an overall amount of 21 runs submitted by 5 different groups. According to the run descriptions submitted by the participants, systems varied from classic IR engines with XML-specific ranking capabilities to highly specialized XQuery engines with full-text extensions. As for efficiency, average running times per topic varied from 91 ms to 17.19 seconds over the entire batch of 568 topics, from 19 ms to 4.72 seconds over the 540 type (A) topics, from 845 ms to 14.58 seconds over the 21 type (B) topics, and from 41 ms to 18.19 seconds over the 7 type (C) topics, respectively. Similarly to the Ad-Hoc Track results, article-only runs generally yielded very good efficiency results, as they clearly constitute an easier retrieval mode, however also at a comparable effectiveness level. Overall effectiveness results were generally comparable to the Ad-hoc Track (albeit using different topics), with the best runs achieving a MAiP value of 0.19 and interpolated (early) precision values of 0.67 at 1% recall (iP[0.01]) and 0.49 at 10% recall (iP[0.10]), respectively. Up to now, none of the systems made use of the XPath-FT-based topic format, which leads to the conclusion that so far only systems previously used in INEX were also used for the Efficiency Track.

In summary, the Efficiency Track will continue in 2009, with a focus on specifically difficult topic types. With the new 2009 INEX collection, based on a 2009 dump of the English Wikipedia, with over 2.5 million articles and billions of elements, we already expect major challenges in the scalability of systems for classic ad-hoc retrieval. Thus, the Efficiency Track will continue to provide an interesting, complementary setting to the Ad-Hoc Track.

5 Entity Ranking Track

In this section, we briefly discuss the Entity Ranking track; further details are in [2].

Search engines supporting typed search, and returning entities instead of just web pages, would enable a simplification of many search tasks. In 2007, INEX has started the XML Entity Ranking track (INEX-XER) to provide a forum where researchers may compare and evaluate techniques for systems that return lists of entities. In entity ranking and entity list completion, the goal is to evaluate how well systems can rank entities in response to a user query; the set of entities to be ranked is assumed to be loosely defined by a generic category, given in the query itself, or by some example entities.

Entity retrieval can be characterized as "typed search." The goal of INEX-XER is to evaluate systems built for returning entities instead of documents. In the specific case of this track, categories assigned to Wikipedia articles are used to define the *entity type* of the results to be retrieved. Topics are composed of a set of keywords, the entity type(s), and, for the list completion task, a set of relevant entity examples.

5.1 Tasks

The two main tasks at INEX-XER 2008 were Entity Ranking (XER) and List Completion (LC). They concern information needs represented as triples of type `<query, category, entity>`. The `category` (that is the entity type) specifies the type of objects to be retrieved. The `query` is a free text description that attempts to capture the information need. The `entity` attribute specifies a set of example instances of the given entity type. ER runs are given as input the `query` and `category` attributes, where LC runs are based on `query` and `entity`. In both cases, the system should return the relevant Wikipedia pages (each page playing the role of an entity surrogate).

Additionally, we performed an Entity Relation Search (ERS) pilot task. The motivation for such task is that searchers may want to know details about previously retrieved entities, and, specifically, their relations with other entities. An example relation search seeks museums in the Netherlands exhibiting Van Gogh's artworks, and the cities where these museums are located. A system needs to first find a number of relevant museums, and then establish correct correspondence between each museum and a city. The ERS task could help explore connections between information retrieval and related fields like information extraction, social network analysis, natural language processing, the semantic web, and question answering. ERS concerns tuples of type `<query, category, entity, relation-query, target-category, target-entity>`. The `query`, `category`, and `entity` are already defined in the entity ranking task. The `relation-query` in form of free text describes the relation between an entity and a target entity. The `target-category` specifies the type of the target entity. `Target-entity` specifies example instances of the target entity type.

5.2 Topics

Topics are composed of a title, that is, a keyword query the user provides to the system, a description and a narrative, that is, natural language explanation of the information need. Additionally, a category field and a set of example entities are contained in the topic. ERS topics also contain fields for the relation-query (i.e., title, description, and narrative), target-category, and example entity pairs.

Participants from eleven institutions have created a small number of (partial) entity examples with corresponding topic text. Candidate entities correspond to the names of articles that loosely belong to categories (for example may be subcategory) in the Wikipedia XML corpus. As a general guideline, the topic title should be type explanatory, i.e., a human assessor should be able to understand from the title what type of entities should be retrieved. Some of the topics have been extended for the ERS pilot task.

5.3 Test Collection

The test collection created during INEX-XER 2008 consists of 35 topics and their assessments in an adapted trec_eval format (adding strata information) for the xinfAP evaluation script. We used as official evaluation measure xinfAP as we performed a stratified sampling on top 100 retrieved entities by each run. The evaluation script is available for download at `http://www.l3s.de/~demartini/XER08/`.

Topics 101-149 are XER topics, in that the participants created these topics specifically for the track, and (almost all) topics have been assessed by the original topic authors. From the originally proposed topics, topics with less than 7 relevant entities and topics with more than 74 relevant entities have been excluded from the test collection (because they would be unstable or incomplete, respectively). Three more topics were dropped, one on request of the topic assessor and two due to unfinished assessments, resulting in a final INEX-XER 2008 test collection consisting of 35 topics with assessments. 23 ERS topics are part of the final collection but relevance judgements for the ERS tasks have not been performed. Together with the 25 XER topics created in 2007, a set of 60 topics is now available for evaluating Entity Retrieval systems.

5.4 Results

Most participants used language model techniques as underlying infrastructure to build their Entity Ranking engines. For both the ER and the LC task the best performing approach uses topic difficulty prediction by means of a four-class classification step [22]. They use features based on the INEX topics definition and on the Wikipedia document collection obtaining 24% improvement over the second best LC approach. Experimental investigation showed that Wikipedia categories helped for easy topics and the link structure helped most for difficult topics. As also shown in last INEX-XER edition (best performing group at INEX-XER 2007), using score propagation techniques provided by PF/Tijah works in the context of ER [20]. The third best performing approach uses categories and links in Wikipedia [16]. They exploit distances between document categories and target categories as well as the link structure for propagating relevance information showing how category information leads to the biggest improvements.

For the LC tasks the same techniques performed well. Additionally, [16] also used relevance feedback techniques using example entities. Here, [13] adapted language models created for expert search to the LC task incorporating category information in the language model also trying to understand category terms in the query text.

6 Interactive Track

In this section, we briefly discuss the Interactive track. For further details, we refer to [19].

6.1 Introduction

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 and again in 2006/2007. Although there has been variations in task content and focus, some fundamental premises has been in force throughout:

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

This has ensured that through a manageable effort, participant institutions have had access to a rich and comparable set of data on user background and user behavior, of sufficient size and level of detail to allow both qualitative and quantitative analysis.

6.2 Task

The document collection used for the 2008 iTrack was the same as was used for most of the other INEX tracks, an extract of 650,000 Wikipedia articles. It was decided to experiment with two categories of search tasks, from each of which the searcher were instructed to select one of three alternative search topics constructed by the track organizers. The two categories of tasks consisted of fact-findings tasks (category 1) and research tasks (category 2).

The tasks were generated to represent information needs believed to be typical for Wikipedia users. The first category, fact-finding, represents search tasks that request specific information for a topic. An example of a fact-finding task is:

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

The second category, research, represents search tasks that require broader information on a topic, which can only be found by collecting information from several documents. An example of a research task is:

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

6.3 Participating Groups

Seven groups initially expressed interest in participating in the track, but in the end only two groups were able to perform experiments.

6.4 System and Experiment Design

The track were run using a java-based retrieval system built within the Daffodil framework [5], which resides on a server at and is maintained by the University of Duisburg. The system returns search results consisting of elements of varying granularity (full Wikipedia articles, sections or sub-sections of articles). Elements are grouped by document in the result list, and up to three high ranking elements are shown per document. When a searcher chooses to examine a document the system shows the entire full text of the document with background highlighting for high ranking elements. In addition it shows a Table of Contents drawn from the XML formatting. From the ToC the searcher can choose individual sections and subsections for closer examination.

Before the experiment, the searchers were given a pre-experiment questionnaire, which collected demographic data. Each search task was preceded with a pre-task questionnaire, to establish searchers' perceptions of the search task. After each task, searchers were asked to fill out a post-task questionnaire, containing questions intended to learn about the searchers' use of and their opinion on various features of the search system, in relation to the task they had just completed. The experiment was closed with a post-experiment questionnaire, which asks searchers' general opinion of the search system. The questionnaire data were logged in a database.

The system was designed to have searchers assess the relevance of each item they looked at. These could be the full articles or article elements. Five different relevance scores were available. The scores expressed two aspects or dimensions in relation to solving the task: 1) How much relevant information does the part of the document contain? It may be highly relevant, partially relevant or not relevant. 2) How much context is needed to understand the element? It may be just right, more or less. All search sessions were logged and saved to a database.

6.5 Findings

Based on the log files, involving 29 test persons, a total of 56 sessions were successfully recorded (14 in Amsterdam and 42 in Oslo). Analysis of the logs and questionnaires is still ongoing; results so far concentrate on searchers' perceptions and performance in relation to the two search tasks.

In general, the results indicate that searchers were more satisfied when completing the research task compared to the fact-finding task. From the questionnaire, we found that test persons regarded the research task easier, were more satisfied with the search result and found more relevant information for the research task. This is plausibly related to the task type, where test persons regard more information as relevant or useful when searching for a more open-ended research task. Fact-finding tasks require a more specific and precise answer, which may diminish the additional value of exploring a wide range of search results. This finding is consistent with the relevance assessment results, from the transaction log, where searchers found more relevant articles and elements when completing the research task compared to the fact-finding task. Also fact-finding sessions resulted in significantly more non-relevant articles than research sessions.

In the log, we see that test persons performed more queries in fact finding session and spent more time to solve research task. In other words, test persons examined the individual article/element more thorough when completing the research tasks. This could be related

to our finding that test persons found more relevant results for the research tasks. This explanation is also supported by the results of our questionnaire stating that the test persons were less certain that they had completed the fact-finding task compared to the research task.

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

7 Link the Wiki Track

In this section, we briefly discuss the Link the Wiki track. A comprehensive discussion can be found in [10].

Automated link discovery in a centralized document repository is a challenging task. Focused link discovery takes the process a step further – the system must link each anchor text in the new document to the best entry point (BEP) in the target document. Incoming links are also focused – new anchors are identified in existing documents and are linked to their respective best entry points in the new document. In a growing collection, such as the Wikipedia, this approach can help keep the link graph up-to-date. This link graph maintenance requirement was motivation for the INEX Link-the-Wiki track.

The Link the Wiki track at INEX 2008 offered two tasks, file-to-file link discovery and anchor-to-bep link discovery. In the file-to-file task 6,600 documents were randomly selected, links removed, and evaluation of discovered links performed against the original collection links. In the anchor-to-bep task 50 topics were nominated by participants. The links discovered by the participants systems were pooled and were exhaustively manually assessed. Runs were evaluated using standard precision and recall measures such as MAP and interpolated precision-recall graphs.

The results suggest that automated link discovery is not a solved problem and that any evaluation of link discovery systems in the Wikipedia must be based on manual assessment, not on the existing links.

7.1 Methodology

The collection used was a dump of the Wikipedia from 2006, consisting of 659,388 articles. A topic in the Link-the-Wiki track was an orphaned article (a de-linked document) and the goal was to extensively link it. In 2007, the Link-the-Wiki track was run at INEX for the first time and only the file-to-file task was run, and only with 90 topics. In 2008, the task was extended to 6600 randomly selected topics. Up to 250 outgoing and 250 incoming links were required per topic. In the new for 2008 anchor-to-bep task, 50 anchors were to be discovered, each anchor having up to 5 links.

A total of 10 groups from 8 different organizations participated in the track. 25 runs were submitted to the file-to-file task. In this task the ground-truth was those links already in the Wikipedia. A total of 31 runs were submitted to the anchor-to-bep task. All runs in the task were pooled for manual assessment. Those links already in the Wikipedia document were also added to the pool. The assessment pool was exhaustively evaluated. Topics contained between 405 and 1,772 links in the pool. A consequence of this approach is that the links already present in the Wikipedia are manually assessed.

A GUI tool was developed to facilitate the efficient manual assessment. Figure 1 shows

Figure 1: Link the Wiki 2008 Assessment Tool

a screenshot of the program. The pool is on the right, the linking document is in the middle and the orphan topic with anchors embedded is on the left. The assessors decided the relevance or nonrelevance of a topic document, or a bunch of links within an anchor, by mouse clicks. The best-entry-point could be positioned in an appropriate position with a double-click; alternatively the link could be declared irrelevant with a right-click. The entire assessment process for a topic took about 4 to 6 hours to finish.

Evaluation of submitted links was performed using two sets of assessments. One set was derived from the existing Wikipedia links. The other was derived from the manual assessments. The evaluation of file-to-file links was based on standard precision/recall measures, treating the submitted list of links as a ranked list and measuring it against the assessment sets. Relevance was binary, either 0 (nonrelevant) or 1 (relevant). For the Anchor-to-BEP evaluation the relevance measure was adapted to include BEP proximity. The proximity of the BEP to a manually designated BEP (measured in character distance) was taken into account to derive the score of the link.

7.2 Results

The main link discovery methods utilized in the runs were based on two approaches: *Anchor Link Analysis* and *Page Name Analysis*. At INEX, the former approach is due to Itakura & Clarke [11] and the latter is due to Geva [7]. Both approaches were first seen at INEX in 2007. The best Anchor Link Analysis run was submitted by the University of Otago [12]. The best Page Name Analysis run was submitted by QUT. Each institute corrected minor coding issues in their algorithms and re-submitted their best run. A third run was generated

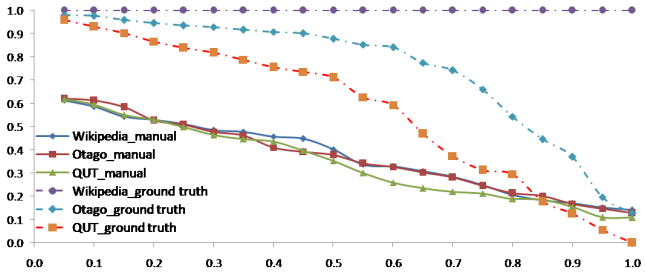


Figure 2: Evaluation against manual and ground-truth assessments

from the Wikipedia by taking the first 50 links in the original Wikipedia document.

Figure 2 is a precision/recall graph showing the results of outgoing file-to-file link analysis against two assessments sets. The ground-truth set was those links present in the Wikipedia documents. The manual set is those links assessed as relevant by the human assessor: six lines are seen, the upper three are the assessment against the Wikipedia ground-truth whereas the lower three are the assessment against the manual assessments. The difference in performance of the three runs is large and significant when compared to the Wikipedia ground-truth, but slight and insignificant when compared to the manual assessments.

7.3 Discussion and Conclusions

Automated links discovery systems based on Anchor Link Analysis can perform near-perfectly when compared to the links already present in the Wikipedia. Those based on Page Name Analysis do not. However, then compared to the manually assessed links, the performance difference is not significant. The gap between linking in Wikipedia and readers' expectation is apparent since the assessors subjectively eliminate *unnecessary links* (e.g. link to year pages). The track has raised the question of how to algorithmically determine the difference between the links in the Wikipedia and those that a human assessor would asses as "relevant." This question will be examined by the track in 2009.

8 XML Mining Track

In this section, we briefly discuss the XML Mining track; a detailed discussion is in [4].

8.1 Aims and tasks

The XML Document Mining track was launched for exploring two main ideas: first identifying key problems for mining semi-structured documents and new challenges of this emerging field and second studying and assessing the potential of machine learning techniques for dealing with generic Machine Learning (ML) tasks in the structured domain i.e. classification and clustering of semi structured documents. This track has run for four editions since INEX 2005, and the fifth edition is currently being launched. Among the many open problems for handling structured data, the track focuses on two generic ML tasks applied to Information Retrieval: while the preceding editions of the track concerned supervised *classification/categorization* and unsupervised *clustering* of independent document, this track is about the classification and the clustering of XML documents organized in a graph of documents. The goal of the track was therefore to explore algorithmic, theoretical and practical issues regarding the classification and clustering of interdependent XML documents.

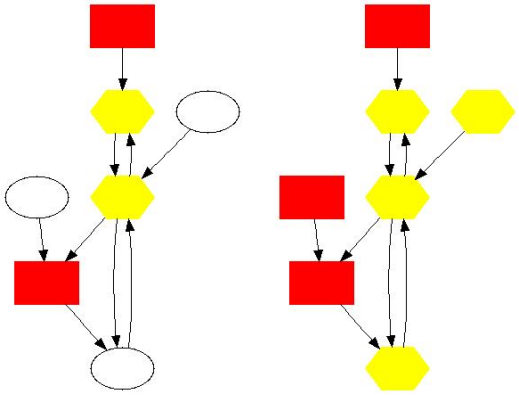
Dealing with XML document collections is a particularly challenging task for ML and IR. XML documents are defined by their logical structure and their content (hence the name semi-structured data). Moreover, in a large majority of cases (Web collections for example), XML documents collections are also structured by links between documents (hyperlinks for example). These links can be of different types and correspond to different information: for example, one collection can provide hierarchical links, hyperlinks, citations, etc. Earlier models developed in the field of XML categorization/clustering simultaneously use the content information and the internal structure of XML documents for a list of models) but they rarely use the external structure of the collection i.e the links between documents.

We have focused on the problem of classification/clustering of XML documents organized in graph. More precisely, this track was composed of:

- a *single label classification* task where the goal was to find the single category of each document. This task consider 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 (see Figure 3).
- a *single label clustering* task where the goal was to associate each document to a single cluster, knowing both the documents and the links between documents (see Figure 4).

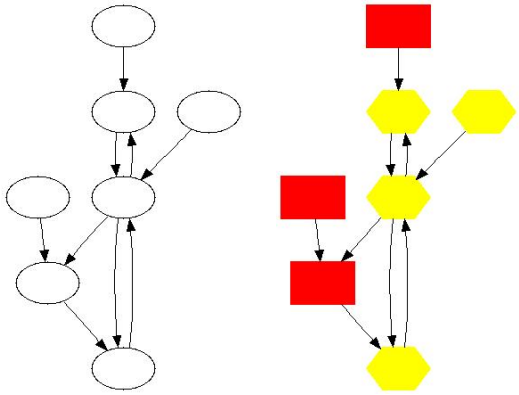
8.2 Collection

The corpus provided is a subset of the *Wikipedia XML Corpus* [3]. We have extracted a set of 114,336 documents and the links between documents. These links corresponds to the links provided by the authors of the Wikipedia articles. Note that we have only kept the links that concern the 114,333 documents of the corpus and we have removed the links that point to other articles. The provided corpus is composed of 636,187 directed links that correspond to hyperlinks between the documents of the corpus. Each document is pointed by 5.5 links on average and provide 5.5 links to other documents. The number of links (in-links and out-links) directly depend on the size of the documents. This means that large documents are more cited than small ones. This characteristic is specific to Wikipedia and does not fit well with Web graph for examples. The global corpus topology is dense: the corpus is composed of one giant component where a large majority of documents are linked to and some very small "islands" of documents that are not linked to this component. The collection contains more than 20,000 possible categories, and one document can belong to many categories. In order



Training set Final labeling

Figure 3: The supervised classification task.



Training set Final labeling

Figure 4: The unsupervised clustering task.

to provide a single label classification/clustering benchmark, we have labeled the documents with a subset of the original Wikipedia categories. These categories have not been chosen randomly in the whole set of categories. We have kept a subset of 15 categories that allow reasonable performances for the supervised classification task using a Naive Bayes classifier. For the categorization task, we have provide the labels of 10% of the documents as a training set. These labels have been chosen randomly amongst the documents of the corpus.

8.3 Evaluation and Results

Each submission has been blinded evaluated by the organizers on the testing corpus. For categorization, we have asked the participants to submit one 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 *recall* that corresponds to the percentage of documents of the category that have been correctly classified.

For the clustering task, the participants have submitted a cluster index for each of the documents of the testing set. We have then evaluated if the obtained clustering corresponds to the real categories of the documents. For each submitted cluster, we have computed a *purity* measure that is a recall of the cluster considering that the cluster belongs to the category of the majority of its documents. We have also used a *micro average purity* and a *macro average purity* in order to summarize the performances of the different models over all the documents and all the clusters. Note that the evaluation of clustering is still an open problem particularly with semi-structured document where clusters can correspond to structural clusters or to thematic clusters. The measures proposed here just gives an idea of how much a model is able to find the 15 categories in an unsupervised way.

Four models have been submitted for the clustering task and five for the supervised classification Detailed results are given in [4]. For classification, the two best models (more than 78% recall) are obtained using classical vector classifiers (SVMs) with an appropriated document representation that mainly only uses the content information and link frequencies. The three other models that better use the graph structure perform between 73.8% and

68.1% in term of recall. For the clustering task, the purity obtained by the best submitted models for 15 clusters is around 50%. Note this purity can directly be compared to the 78% recall obtained by the supervised methods showing that supervision improves unsupervised learning by 28%.

9 Envoi

This complete our walk-through of the seven tracks of INEX 2008. 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 2008 are being published in the Springer LNCS series [8]. This volume contains both the track overview papers, as well as the papers of the participating groups. The main result of INEX 2008, however, is a great number of test collections that can be used for future experiments.

INEX 2009 will see some exciting changes. First and foremost is the creation of a new collection, again based on the Wikipedia but a 2009 crawl containing over 2.5 million articles (making it four times larger than the current collection). Most of the track will continue, with similar tasks on the new collection, or entirely new tasks that address other aspects of focused retrieval.

References

[1] M. S. Ali, M. P. Consens, G. Kazai, and M. Lalmas. Structural relevance: a common basis for the evaluation of structured document retrieval. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1153–1162, New York, NY, USA, 2008. ACM.

[2] G. Demartini, A. P. de Vries, T. Iofciu, and J. Zhu. Overview of the INEX 2008 entity ranking track. In Geva et al. [8].

[3] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40(1):64–69, June 2006.

[4] L. Denoyer and P. Gallinari. Overview of the INEX 2008 XML mining track. In Geva et al. [8].

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

[6] N. Fuhr, M. Lalmas, A. Trotman, and J. Kamps, editors. *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, number 4862 in LNCS. Springer Verlag, Berlin, Heidelberg, 2008.

[7] S. Geva. GPX: Ad-hoc queries and automated link discovery in the wikipedia. In Fuhr et al. [6], pages 404–416.

[8] S. Geva, J. Kamps, and A. Trotman, editors. *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, LNCS. Springer Verlag, Berlin, Heidelberg, 2009.

[9] N. Gövert and G. Kazai. Overview of the Initiative for the Evaluation of XML retrieval (INEX) 2002. In N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, editors, *INEX Workshop*, pages 1–17, 2002.

[10] W.-C. Huang, S. Geva, and A. Trotman. Overview of the INEX 2008 link the wiki track. In Geva et al. [8].

[11] K. Itakura and C. Clarke. University of Waterloo at INEX 2007: Ad hoc and link-the-wiki tracks. In Fuhr et al. [6], pages 380–387.

[12] D. Jenkinson, K.-C. Leung, and A. Trotman. Wikisearching and wikilinking. In S. Geva, J. Kamps, and A. Trotman, editors, *INEX 2008 Workshop Pre-proceedings*, pages 330–344, 2008.

[13] J. Jiang, W. Lu, X. Rong, and Y. Gao. Adapting Expert Search Models to Rank Entities. In Geva et al. [8].

[14] J. Kamps, S. Geva, A. Trotman, A. Woodley, and M. Koolen. Overview of the INEX 2008 ad hoc track. In Geva et al. [8].

[15] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. INEX 2007 Evaluation Measures. In Fuhr et al. [6], pages 24–33.

[16] R. Kaptein and J. Kamps. Finding Entities in Wikipedia using Links and Categories. In Geva et al. [8].

[17] G. Kazai, A. Doucet, and M. Landoni. Overview of the INEX 2008 book track. In Geva et al. [8].

[18] G. Kazai, N. Milic-Frayling, and J. Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *SIGIR '09: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 2009.

[19] N. Pharo, R. Nordlie, and K. N. Fachry. Overview of the INEX 2008 interactive track. In Geva et al. [8].

[20] H. Rode, D. Hiemstra, A. P. de Vries, and P. Serdyukov. Efficient XML and Entity Retrieval with PF/Tijah: CWI and University of Twente at INEX'08. In Geva et al. [8].

[21] M. Theobald and R. Schenkel. Overview of the INEX 2008 efficiency track. In Geva et al. [8].

[22] A.-M. Vercoustre, J. Pehcevski, and V. Naumovski. Topic Difficulty Prediction in Entity Ranking. In Geva et al. [8].