Report on INEX 2009

T. Beckers P. Bellot G. Demartini L. Denoyer C.M. De Vries
A. Doucet K.N. Fachry N. Fuhr P. Gallinari S. Geva
W.-C. Huang T. Iofciu J. Kamps G. Kazai M. Koolen
S. Kutty M. Landoni M. Lehtonen V. Moriceau R. Nayak
R. Nordlie N. Pharo E. SanJuan R. Schenkel X. Tannier
M. Theobald J.A. Thom A. Trotman A.P. de Vries

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 2009 evaluation campaign, which consisted of a wide range of tracks: Ad hoc, Book, Efficiency, Entity Ranking, Interactive, QA, Link the Wiki, and XML Mining. INEX in running entirely on volunteer effort by the IR research community: anyone with an idea and some time to spend, can have a major impact!

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 2009 evaluation campaign consisted of a wide range of tracks:

Ad hoc Track Investigating the effectiveness of XML-IR and Passage Retrieval for four ad hoc retrieval tasks (Thorough, 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.

Question Answering Track Investigating how technology for accessing semi-structured data can be used to address interrogative information needs.

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 the rest of this paper, we discuss the aims and results of the INEX 2009 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 QA track (Section 7), the Link the Wiki track (Section 8), and the XML Mining track (Section 9).

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 [3].

2.1 Aims and Tasks

The Ad Hoc Track at INEX studies the adhoc retrieval of XML elements or passages. 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 2009 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 2009 Ad Hoc Track featured four tasks: For the *Thorough Task* a ranked-list of results (elements or passages) by estimated relevance must be returned. It is evaluated by mean average interpolated precision. For the *Focused Task* a ranked-list of non-overlapping results (elements or passages) must be returned. It is evaluated at early precision. 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

Starting in 2009, INEX uses a new document collection based on the 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 [16, 14], explicitly labeling more than 5,800 classes of entities like persons, movies, cities, and many more.

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 10, 2009. At this time 68 topics had been fully assessed.

The main INEX 2009 test-collection consists the 68 passage-level judged topics, and the specific measures to evaluate the four 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.

2.3 Results

We received 172 submissions from 19 participating groups, which are discussed in detail in the individual papers in [4]. There were three main research questions underlying the Ad Hoc Track. The first main research question was the impact of the new collection—four times the size, with longer articles, and additional semantic markup—on focused retrieval. That is, what is the impact of collection size? What is the impact of document length, and hence the complexity of the XML structure in the DOM tree? We saw that the collection's size had little impact, but that the relevant articles were much longer (a mean length 3,030 in 2008 and 5,775 in 2009, a 52% increase), leading to a lower fraction of highlighted text per article (a mean of 58% in 2008 and 33% in 2009). This also reduced the correlation with article retrieval, e.g., from 79% for the "in context" tasks in 2008 to 51–58% in 2009. Obviously locating the "right" articles is important, but not enough to score well on the focused retrieval tasks.

The second main research question was the impact of more verbose queries—using either the XML structure, or using multi-word phrases. That is, what is the impact of semantic annotation on both the submitted queries, and their retrieval effectiveness? What is the impact of explicitly annotated multi-word phrases? We found that for all tasks the best scoring runs used the CO query but some CAS runs were in the top 10 for all four tasks. Part of the explanation may be in the low number of CAS submissions (40) in comparison with the number of CO submissions (117). Only 50 of the 68 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. The YAGO tags potentially expressing an information need naturally in terms of structural constraints, were popular: 36 CAS queries used them (21 of them judged). Over the 50 non-trivial CAS queries, most groups had a better performing

run using the CO query, although the structural hints were useful for some individual topics and did lead to a precision gain for some participants.

There were also few submissions using the explicitly annotated phrases of the phrase query: ten in total. Phrase query runs were competitive with several of them in the overall top 10 results, but the impact of the phrases seemed marginal. Note that the exact same terms were present in the CO query, and the only difference was the phrase annotation.

The third main research question is that of the value of the internal document structure (mark-up) for retrieving relevant information. That is, does the document structure help to identify where the relevant information is within a document? The number submissions using File-Offset-Length (FOL) passages and range of elements was low. Thirteen submissions used ranges of elements or FOL passage results, whereas 144 submissions used element results. Some of the non-elemental submissions were competitive, but these runs were typically based on article or element runs. The outcome broadly confirms earlier results that the document structure helps select the best passages to retrieve.

For all these research questions we hope and expect that the resulting test collection will prove its value in future use. After all, the main aim of the INEX initiative is to create bench-mark test-collections for the evaluation of structured retrieval approaches.

2.4 Outlook

Plan for INEX 2010 are currently under discussion. One of the main research questions will be the impact of resource-limitations and effort within focused retrieval. This could be studied using new measures that take the reading effort into account, by imposing limitations on the length of results (e.g., inspired by what fits on a screen), or by explicitly framing focused retrieval as a snippet or teaser generating task.

3 Book Track

In this section, we briefly discuss the Book Track. For further details, please refer to [7].

3.1 Goals and Setup

The aim of the Book Track is to evaluate approaches for supporting users in reading, searching, and navigating the full texts of digitized books. In 2009, the track focused on four evaluation tasks:

- The *Book Retrieval* (BR) task, framed within the user task of building a reading list for a given topic of interest, e.g., for a Wikipedia article, aimed at comparing traditional document retrieval methods with domain-specific techniques that exploit book-specific features, such as the back of book index or library catalogue information,
- The *Focused Book Search* (FBS) task aimed to test the value of applying focused retrieval approaches to books, where users expect to be pointed directly to relevant book parts,
- The *Structure Extraction* (SE) task aimed at evaluating automatic techniques for deriving structure from OCR and layout information for building hyperlinked table of contents, and,

- The *Active Reading task* (ART) aimed to explore suitable user interfaces for eBooks enabling reading, annotation, review, and summary across multiple books.

A total of 84 organisations registered for the track, of which 16 took part actively throughout the year, contributing topics, runs, or relevance judgements to the test collection.

3.2 Test Collection

The test collection consists of 50,239 digitized out-of-copyright books (totaling 400GB), including history books, biographies, literary studies, religious texts and teachings, encyclopedias, 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. Both the BR and FBS 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 1000 books for which JPEG page images and the original OCR files (in DjVu XML, with only page level structure) were distributed to participants.

Participants created 16 new topics this year, containing 37 topic aspects. The aspects were used in the FBS task, while the full topics in the BR task.

Relevance assessments were collected using the 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 [8]. In 2009, assessments were gathered through a series of 'Read and Play' games: In game 1, participants had the task of finding books relevant to a given topic and then ranking the top 10 most relevant books. In game 2, their task was to find pages relevant to a given topic aspect inside the books that were selected in game 1. Finally, in game 3, their task was to review the pages that were judged in game 2.

The games run for three weeks with weekly prizes of $50 worth of Amazon gift vouchers, shared between the top three scorers, proportionate to their scores. In total, 4,403 book level relevance judgements were contributed by 9 assessors in game 1, but only 235 pages were judged by 2 assessors in games 2 and 3 (although 1,435 pages were also judged in game 1, these were only judged with respect to the overall topic, and not the specific topic aspect).

3.3 Results

3.3.1 Book Retrieval and Focused Book Search Tasks

Due to the limited amount of page level judgements, results were only published for the BR task. 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, 21 runs were submitted by 4 groups. Participants experimented with various techniques, e.g., using table of contents or the back-of-book index as well as traditional document retrieval methods. The best performing run (MAP=0.3731) was submitted by the University of Amsterdam team, which ranked books (treated as documents) using Indri and applied query expansion with 50 terms from the top 10 results. Although performance improved across the board compared with last year (best MAP in 2008 was 0.1056),

traditional document retrieval approaches, i.e., not specific to books, still dominated the top ranks. This suggests that there is still plenty to be done in discovering suitable ranking strategies for books.

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, using a structure labeling tool developed at the University of Caen. Performance was evaluated using recall/precision like measures at different structural levels (i.e., different depths in the ToC). 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.

8 runs were submitted by 4 groups. The best performance (F=41.51%, calculated as the harmonic mean of precision and recall) was obtained by the Microsoft Development Center Serbia team, who extracted ToCs by first recognizing the page(s) of a book that contained the printed ToC. Interestingly, they gave the best result last year too on a test set of only 100 books (F=53.47%).

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. The task has only attracted 2 groups, neither of whom submitted any results at the time of writing.

3.4 Conclusions and Outlook

The Book Track in 2009 has attracted a lot of interest and has grown considerably from the previous years. However, active participation remained a challenge for most. A reason may be the high initial setup costs (e.g., building infrastructure to search books). At the same time, the Structure Extraction task has been met with great interest and created a specialist community at ICDAR'09. The search tasks, although explored real-world scenarios around providing reference sources for Wikipedia articles, were only tackled by a small set of groups. Since the evaluation of the BR and FBS tasks requires a great deal of effort, e.g., developing the relevance assessment system, designing the games, and collecting the judgements, we will be re-thinking the setup of these tasks for INEX 2010. For example, we plan to concentrate on focused (narrow) topics for which only few pages in the corpus may be relevant. In addition, to improve the quality of the test topics, we will look for ways to automate topic creation. To provide real value in improving the test corpus itself, we plan to run the SE task with the goal to convert the current corpus to an XML format that contains rich structural and semantic markup, which then can be used in subsequent INEX competitions. To attract participants, we plan to run ART either at a separate forum or combined with the interactive track.

4 Efficiency Track

In this section, we discuss the goals, general setup and results of the Efficiency Track. For further details, we refer to [15].

4.1 Overview

The Efficiency Track was run for the second time in 2009, with its first incarnation at INEX 2008 [17]. It was intended to provide a common forum for the evaluation of both the effectiveness and efficiency of XML ranked retrieval approaches on *real data* and *real queries*. 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).

The Efficiency Track used the INEX-Wikipedia collection (available from `http://www.mpi-inf.mpg.de/departments/d5/software/inex/`) that has been introduced in 2009, an XML version of English Wikipedia articles with semantic annotations. With almost 2.7 million articles, more than a billion elements and an uncompressed size of approximately 50 GB, this collection is significantly larger than the old Wikipedia collection used in previous years (and for last year's Efficiency track). The collection has an irregular structure with many deeply nested paths, which turned out to be challenging for most systems.

4.2 Topics, Tasks and Metrics

Topics. One of the main goals of the Efficiency Track was covering a broader range of query types than the typical AdHoc queries with only a few keywords and hardly any structural conditions. Thus, in addition to the existing 115 topics from the AdHoc track (labeled *type A topics*), the track also considered 115 topics (*type B topics*) generated by running Rocchio-based blind feedback on the results of the article-only AdHoc reference run. These keyword-only topics were intended to simulate high-dimensional query expansion settings with up to 101 keywords, which cannot be evaluated in a conjunctive manner and posed a major challenge to any kind of search engine. The track should also contain *type C topics*, high-dimensional, structure-oriented retrieval settings over a DB-style set of content-and-structure queries with deeply nested structure but only a few keyword conditions, but we did not get any proposals for type (C) topics by the track participants.

Tasks. The Efficiency Track particularly encouraged the use of top-k style query engines and thus asked participants to create runs with the top-15, top-150, and top-1500 results, and to report runtimes and I/O costs. In addition to two standard tasks from the AdHoc track, i.e. Focused and Thorough, the track introduced two extra tasks: The *Article Task* (which asked for results at the article level only) and the *Budget-Constrained Task*. In the latter task, which was introduced in 2009, systems should retrieve results within a fixed budget of runtime, simulating interactive retrieval situations. Standard top-k algorithms cannot easily be used with such a constraint or may return arbitrarily bad results. However, we did not get any submissions for this task, probably because it required specific modifications to the systems which was too much effort for the participants.

Metrics. The Efficiency Track applied the same metrics and tools used in the Ad-Hoc track to assess result quality, and additionally considered runtime as an equally important metrics.

4.3 Results and Conclusions

The track received 68 runs submitted by 4 participating groups using 5 different systems, which is comparable to 2008. This included two element retrieval systems (one of which was implemented as a distributed system), one XPath FullText system, and two article retrieval systems. Regarding efficiency, average running times per topic varied from 8.8ms to 50 seconds for the type A topics and from 367.4 ms to 250 seconds for the type B topics. While absolute runtimes across systems are hardly comparable due to differences in the hardware setup, it became evident that the dominant part of retrieval time was spent in I/O activity, so improving or reducing I/O access could be a promising way to improve efficiency. Actually, one of the participating systems already kept the index completely in memory, requiring additional I/O only to postprocess result documents, which could be stored in a sufficiently large memory as well. Result quality was comparable to the AdHoc Track on the type A topics, with the best runs achieving a MAiP value of 0.301 and an iP[0.01] of 0.589. Result quality on the type B topics was generally slightly worse compared to the type A topics, which was probably caused by the extreme query expansion.

The 2009 Efficiency Track has demonstrated that a number of systems are able to achieve very good result quality within very short processing times that allow for interactive retrieval. A future Efficiency Track therefore needs to introduce new challenges, which could come from a much larger collection (such as ClueWeb), from more complex queries with more structural conditions, or from a combination of both.

5 Entity Ranking Track

In this section, we will briefly discuss the INEX 2009 XML Entity Ranking track. Further details are in [1].

5.1 Overview

Search engines supporting typed search, and returning entities instead of just web pages, would facilitate 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. As compared to previous years, INEX-XER 2009 used a new Wikipedia collection which is more recent, bigger in size, and contains entity annotations in the text. 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.

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.

A set of 60 topics has been selected from those developed by INEX-XER participants during the past years. Candidate entities correspond to the names of articles that loosely belong to categories (or subcategories) 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.

5.3 Tasks

The two tasks at INEX-XER 2009 were Entity Ranking (ER) and List Completion (LC). They concern information needs represented as triples of type `<query, category, entity>`. The `category` (that is 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).

5.4 The INEX-XER 2009 test collection

The final test collection created during INEX-XER 2009 consists of 55 topics and their assessments, which are in an adapted trec_eval format (adding strata information) for the xinfAP [19] evaluation script. The INEX-XER 2009 test collection is available for download at `http://www.l3s.de/~demartini/XER09/`. We used as official evaluation measure xinfAP as we performed a stratified sampling on top 100 retrieved entities by each run. Compared to last year, a less aggressive sampling strategy was used in order to allow the judgment of more topics (49 in 2008 versus 60 in 2009) with the same resources.

Topics were created by previous years participants specifically for the track, and have been re-assessed this year on the new Wikipedia collection. Consistently with last year, 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 for the LC task, as the example entities were not relevant anymore in the new collection. The final INEX-XER 2009 test collection consists of 55 topics with assessments for ER and 52 for LC.

Additionally, INEX-XER 2009 created a set of not-an-entity annotations for Wikipedia pages such as list-of or disambiguation pages. Such annotations do not influence the evaluation of XER systems as not-an-entity pages are considered as non-relevant entities. They may however be useful as training data, for example, to build classifiers for entity/non-entity pages.

Together with the collections created in 2007 and in 2008, a set of 55 topics is now available for evaluating Entity Retrieval systems on two different Wikipedia collections.

5.5 Results

We noticed that a common behavior of participants this year was to identify entity mentions in the text of Wikipedia articles, passages, or queries. They then applied different techniques (e.g., detect entity relations, exploit category information) to produce a ranked list of Wikipedia articles that represents retrieved entities. The best performing approach exploited a probabilistic framework ranking entities using similarity between probability distributions obtaining an estimated Average Precision of 0.52 for both ER and LC tasks.

This has been the last year for INEX-XER which has built, over three years, a set of 55 ER topics with relevance assessments over two different document collections. Moreover, we have observed, over three years, an improvement in term of effectiveness and more advanced techniques being used by Entity Retrieval systems participating to the track. The original research agenda can also be taken further by the Web Entity Ranking Task introduced at TREC 2009.

6 Interactive Track

In this section, we will briefly discuss the Interactive Track. Further details are in [13].

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, in 2006/2007 and again in 2009. 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; and
- 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 Aims and Tasks

For the 2009 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 categories of tasks were intended to reflect the most common purposes a searcher would have for visiting a database of primarily bibliographic data, a *broad, explorative task* (Category I) and a narrower, more *specific, purpose-driven task* (Category II). The *self-selected task* (Category III) was intended to force the searcher

to make a more quality-driven task than the two others. The experiment was designed to let searcher assess the relevance of books, and they could also simulate the purchase of a book by adding it to a basket.

The broad tasks were designed to investigate thematic exploration which will give us data on query development, metadata type preference and navigation patterns. An example of a broad task is:

> *You are considering to start studying sociology. In order to prepare for the course you would like to get acquainted with some good and recent introductory texts within the field as well as some of its classics.*

The narrow tasks were designed to represent relatively narrow topical queries where the purpose is to allow us to study the basis for relevance decisions and compare the searchers' preference of different document representations. An example of a narrow task is:

> *Find trustworthy books discussing the conspiracy theories which developed after the 9/11 terrorist attacks in New York.*

6.3 Test Collection

The document collection used for the 2009 iTrack differed from the collections used in previous years. A crawl of 2 million records from the book database of the online bookseller Amazon.com was 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.

The search system was developed at the University of Duisburg-Essen. It is based on *Daffodil* [2] and partially on *ezDL1* while the retrieval component was implemented using Apache Solr.

6.4 Results

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.

For this track, 41 volunteers were recruited mostly from students of computer science, cognitive and communication science, library science and related fields. 24 of them were male and 17 of them female. Their average age was about 28. On average, they have used the Internet for 9.5 years. All participants had experience with web search engines, searching in digital libraries or digital bookstores.

The participants were given the possibility to express positive as well as negative general comments in the questionnaires. The user interface was praised because it is well arranged and everything fits on a single screen without the need to scroll up or down. The inclusion of another document aspect, namely the reviews, was also pointed out positively by the participants. Technical problems of the search system and sometimes useless related term suggestions due to topical limitations of the data source were points of criticism. Participants also missed highlighting of query terms and filtering options for the result list. Also, the

heterogeneity of the data was disliked, that is, some books have extensive metadata while other have just scarce descriptions.

The searchers were also asked to indicate on a five point scale how useful (5 for very useful) different types of metadata were for solving their search tasks. We found that document titles, publishers' book descriptions and reviews (by users) were the three most popular metadata fields. It is also worth noting that the searchers found keywords (from Amazon) to be more useful than the user-created tags. It seems that searchers put more trust in authoritative sources that use a controlled vocabulary than users' idiosyncratic tagging.

Log analysis shows that the average length of a session decreased from 829s (Category I) and 725s (Category II) to 622s (Category III). The avgerage duration of a search (that is, the time between two queries) was 99s (Category I), 92s (Category II), and 83s (Category III). For more explorative working tasks, longer and more searches were performed.

At the end of a session the participants had 6.61 (I), 4.96 (II), and 1.15 (III) books respectively in their basket. The participants collected more books for the broad tasks than for the narrow tasks. The average query length (number of terms in the simple query field) was 1.99 (I), 2.46 (II), and 2.21 (III). The broadest tasks resulted in the shortest average query length.

6.5 Outlook

The track generated interesting data for further analysis. In particular on the patterns of interaction for the different categories of tasks. Also we would like perform further analysis of the use of different metadata types. The 2009 Track can be considered as an extensive pilot experiment with access to hetergeneous data. Based on our experiences, a larger scale experiment will run as part of INEX 2010.

7 Question Answering Track

In this section, we will briefly discuss the new methods that we propose in order to compare QA and focused IR systems when a short answer is required, as well as QA and summarization systems when aggregated results are expected. Further details are in [10].

7.1 Motivation

Evaluation campaigns for Question-Answering (QA) systems (for example, TREC, CLEF, etc.) aim at evaluating systems that retrieve precise answers rather than documents in response to a question. The question test sets are generally composed of factoid questions (questions which require a single precise answer to be found in the document collection) and sometimes of more complex questions (list, "how" and "why" questions). Complex questions introduced in these campaigns expect a phrase (or a sentence) as an answer for "how" and "why" questions and a set of distinct short items to be found in the whole collection for list questions. Best systems reach 70% of correct answers for factoid questions but only 40% for list questions [18].

The INEX 2009 QA@INEX track aims to compare the performance of QA, XML/passage retrieval and automatic summarization systems on an encyclopedic resource (Wikipedia). The track considers two types of questions: factoid questions and more complex questions whose answers require the aggregation of several passages.

7.2 Proposed evaluation of Short Answers

Short parts of text (one or a very few words) are the most usual way to answer factual questions in so-called question-answering systems. These questions are the classical ("basic") set that state-of-the-art QA systems are expected to answer quite well.

For QA@INEX, participants need to provide a small ordered set (10) of non overlapping XML elements or passages that contains a possible answer to the question and for each element or passage, the position of the answer found by the system in the passage.

Answers are evaluated by computing their distance to the real answer. The evaluation methodology differs from traditional QA campaigns, where a short answer must be provided besides the supporting passage. This is a major difference in terms of metrics used to rank the participating systems. We then assess an answer not through the full/incomplete paradigm, but rather by *the distance between the indicated answer entry point and the real one.*

This new way to evaluate QA systems has an interesting side effect: it allows focused IR systems to participate in this task using the same evaluation, even if they are unable to extract a short answer or if they have very basic techniques to do so. These systems may simply provide the most relevant and short extracted passages they retrieve, and set an entry point wherever they can in this text. For the first time, QA, XML retrieval and other focused IR systems can participate to the same campaign.

7.3 Proposed evaluation of long answers

INEX has a thorough experience in evaluating focused retrieval systems, however the QA "long answer" subtask is new in this context. The idea here is to propose a common task that can be processed by three different kinds of systems: QA systems providing list of answers, automatic summarization systems by extraction and focused IR systems.

In this QA task, answers have to be built by aggregation of several passages from different documents on the Wikipedia. The maximal length of the abstract being fixed, the systems have to make a selection of the most relevant information. Standard QA systems can produce a list of answers with their support passages. Focused IR systems can return the list of the most relevant XML elements. Note that in this task, IR systems that only retrieve entire documents are strongly handicapped, except if they are combined with automatic summarization systems that build an abstract of the most relevant documents.

Two main qualities of the resulting abstracts need to be evaluated: readability and informative content. The *readability* and coherence of the resulting abstracts are evaluated according to "the last point of interest" in the answer. It requires a human evaluation where the assessor indicates where he misses the point of the answers because of highly incoherent grammatical structures, unsolved anaphora, or redundant passages. The *informative content* of the answer has to be evaluated according to the way they overlap with relevant passages following the experiment in [9] done on TAC 2008 automatic summarization evaluation data and using the FRESA perl package by J-M Torres Moreno (University of Avignon). This allows to directly evaluate summaries based on a selection of relevant passages.

7.4 General time line and available resources

2009 has been devoted to fix the tasks and the overall evaluation methodology based on the corpus, topics and qrels from INEX 2009 ad-hoc track. A first list of questions has been

released for test. They all deal with 2009 INEX topics, hence answers are part of ad-hoc relevant passages. We have annotated correct answers among passages for a subset of 100 short type answers and all long type ones. These resources along with the evaluation programs written in Perl are available for active participants. In order to facilitate submissions from Focused IR systems, a Perl program that converts a run in INEX ad-hoc submission FOL format into QA format is also available. In 2010, we will use the same corpus from ad-hoc INEX 2009 task.

8 Link the Wiki Track

In this section, we will briefly discuss INEX 2009 Link the Wiki Track. Further details are in [5].

8.1 Overview

In the third year of the Link the Wiki track the focus was on focused link discovery.The participants were encouraged to utilize different technologies to identify anchors and best-entry-points (BEPs) for each link. The INEX 2009 Wikipedia collection was used [3]. 5,000 topics were randomly selected for the file-to-file task and 33 topics were nominated by participants for the anchor-to-BEP task. The assessment tool and the evaluation tool were revised to improve efficiency. The file-to-file runs were evaluated against the Wikipedia ground-truth, while the anchor-to-BEP runs were evaluated against a manual assessed set as well as the Wikipedia itself. A new focus-based metric was introduced. Two new tasks involved linking the Te Ara Encyclopedia. There are no hyperlinks in Te Ara and so the first task was to cross link the entire collection. The second Te Ara task was to link to the INEX 2009 English Wikipedia collection. Te Ara runs have not yet been evaluated. Results suggest that existing automatic link discovery algorithms for the Wikipedia can produce better quality links than those presently in the Wikipedia!

8.2 Tasks and Submissions

For the Wikipedia file-to-file task, 5,000 articles were randomly selected, but filtered by document size and the number of out-going links to ensure suitability of the documents to the task. For the Wikipedia anchor-to-BEP task, a total of 33 topics were nominated by participants and were manually assessed. The INEX tasks assume link-discovery is a recommendation task. The system produces a ranked list of source anchors, and for each anchor, a set of target best entry points (BEPs) from which a user can choose. Evaluation is, consequently, based on a combination of a rank-based score for anchors and a set-based score for targets. For a description of the metrics see the INEX pre-proceedings. Six groups participated in the Wikipedia tasks with 29 runs submitted. All Wikipedia runs were evaluated against the Wikipedia ground truth. The anchor-to-BEP runs were additionally evaluated in several different ways including: anchor-to-file (conventional linking) and anchor-to-bep (focused link discovery).

The Te Ara Encyclopedia contains 3,179 articles and is about 50MB of XML. There are no hyperlinks in the collection as it has never been linked (not even manually). Page names are not as indicative of content as they are in the Wikipedia because Te Ara is divided into

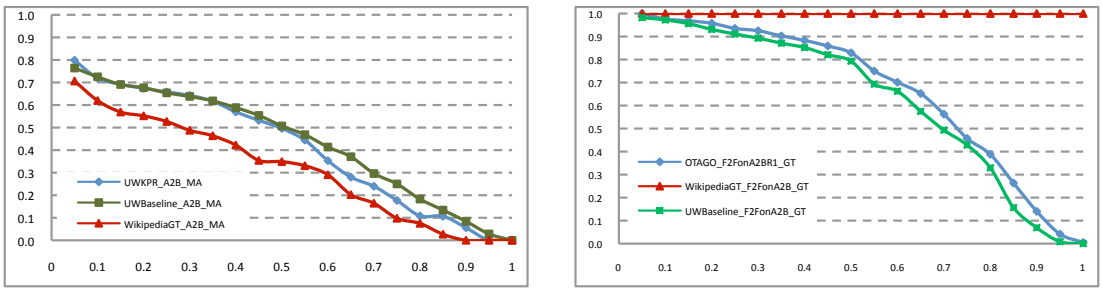

Figure 1: File-to-file evaluation using automatic qrels (left) and Anchor-to-BEP evaluation using manual qrels (right).

stories each of which contains several pages chained together. The task was to link the entire encyclopedia and new approaches were expected because neither link mining nor page name matching were likely to be effective. Two groups participated in the Te Ara tasks with 7 runs submitted.

A run validation tool was introduced in 2009. This tool displays the source document with anchors highlighted, target documents and a list of the links in a run. Runs were pooled and assessed to completion. The pools contained about 1,000 outgoing and 900 incoming links. An assessment tool was supplied to facilitate manual assessment. Assessment is a slow and laborious task; logs suggest it took about 4 hours per topic.

8.3 Most Effective Approaches

Figure 1 shows the evaluation of anchor-to-BEP and file-to-file submissions done using the Wikipedia ground truth and the manual qrels. The links in the Wikipedia were generated through careful construction by a user, or automatically by matching page names, either way such links are relatively easy to find. According to our experiment, document-level link discovery engines are good at exhibiting high precision levels at most points of recall and systems are scalable. However, the automatic evaluation result could be either optimistic while article title links are always the case, e.g. year, or pessimistic if the relevant anchor links were not specified in the collection and are seen as non-relevant in evaluation. Therefore, the manual evaluation process was undertaken. Based on the link mining approach, an algorithm using the K-L divergence of the PageRank and Topical PageRank for frequent phrases in the source documents outperformed the Wikipedia itself. The detailed description of algorithms can be found in [5].

8.4 Conclusions and Discussion

In 2009 the focus shifted from file-to-file to anchor-to-BEP linking. A validation tools was introduced. Metrics were adapted to the new task. High quality Wikipedia link discovery can be achieved using existing approaches. Manual assessment suggests that it is possible to automatically identify higher quality links than those in the Wikipedia. There does, however, remain room for improvement.

9 XML Mining Track

In this section, we will briefly discuss the XML Mining track. Further details are in [12].

9.1 Aims and Tasks

Mining of XML documents has been perceived as an effective solution to improve XML data management by facilitating better information retrieval, data indexing, data integration and query processing [11]. Due to the inherent flexibility of XML, in both structure and semantics, mining useful information from XML documents is faced with new challenges as well as benefits. The aims of the INEX 2009 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 2009 XML Mining track included two tasks: (1) unsupervised clustering task and (2) semi-supervised classification task. The clustering task requires the participants to group the documents into clusters without any knowledge of cluster labels using an unsupervised learning algorithm. On the other hand, the classification task requires the participants to label the documents in the dataset into known classes using a supervised learning algorithm and a training set. The clustering task in INEX 2009 was launched to explicitly test the cluster hypothesis [6], which states that documents that cluster together have a similar relevance to a given query. It uses manual query assessments from the INEX 2009 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. The classification task in INEX 2009 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 track was composed of:

- 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.
- 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.

9.2 Test Collection

The clustering task collection consists of about 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). 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. There are a total of 1,970,515 terms in this collection after applying stemming, stop-word removal and elimination of terms

that occur in a single document. There are a total of 5,213 unique entity tags and 110,766,016 unique links in the collection. There are a total of 348,552 categories that contain all documents except for a 118,685 document subset containing no category information. These categories are derived by using the YAGO ontology [16] and appear to follow a power law distribution.

A subset of collection containing about 54,889 documents was also used in the clustering task for teams that were unable to process such a large data collection. The set of 54,889 documents and the links between these documents are also used as the test collection in classification task. These links correspond to the links provided by the authors of the Wikipedia articles. There were a total of 186,723 unique terms in this subset collection. The documents belong to 39 categories that correspond to 39 Wikipedia portals. We have provided the labels of 20% of the documents as the training set. The corpus is composed of 4,554,203 directed links that correspond to hyperlinks between the documents of the corpus. Each document is concerned by 84.1 links on average.

9.3 Evaluations and Results

Participants were asked to submit multiple clustering solutions containing different numbers of clusters such as 100, 500, 1,000, 2,500, 5,000 and 10,000. 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 utilise 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 a purity measure that is a recall of the cluster considering that the cluster belongs to the category of the majority of its documents. It is important to note that the class labels are not used in the process of clustering, but only for the purpose of evaluation of the clustering results.

New in the INEX 2009 XML Mining track, the clustering solutions are also evaluated to determine the quality of cluster relative to the optimal collection selection goal. The task is to evaluate how well the clustering solutions are able to group a large document collection in an optimal manner in order to satisfy queries while minimising the search space. A total of 68 topics used in Ad Hoc track were utilised to evaluate the quality of clusters generated on the full set of collection of about 2.7 million documents. A total of 52 topics were used to evaluate the quality of clusters generated on the subset of collection of about 50,000 documents. A total number of 4,858 documents were found relevant by the manual assessors for the 68 topics. The Normalised 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 six research teams have participated in the INEX 2009 clustering task. Two of them submitted the results for the subset data only. Detailed results are given in [12]. As expected as the cluster numbers dividing the data set increases, performance of all the teams based on the Purity score increases 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 (previously unseen) 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. We have also computed an Average precision (APR) score over the list of categories returned for each document. It measures the ability of a system to rank correctly the relevant categories. Five different teams had participated to the task. All teams except one team were able to achieve micro and macro F1 between 50-60%. The APR score was obtained in the range of 68-72% by all teams. Detailed results are given in [12].

10 Envoi

This complete our walk-through of the seven tracks of INEX 2009. 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 2009 are being published in the Springer LNCS series [4]. This volume contains both the track overview papers, as well as the papers of the participating groups. The main result of INEX 2009, however, is a great number of test collections that can be used for future experiments.

INEX 2010 will see some exciting changes. At INEX 2009, most tracks continued earlier tasks to study the effect of the changes in the document collection (in a sense a longitudinal study over three years of Wikipedia). However, 2010 will be a year of innovation with a range of interesting news tasks, and the discussion is still ongoing at the time of writing. As INEX is an entirely volunteer run initiative, anyone with an interesting idea and some time to spend can make an impact!

References

[1] G. Demartini, T. Iofciu, and A. P. de Vries. Overview of the INEX 2009 entity ranking track. In Geva et al. [4].

[2] 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.

[3] S. Geva, J. Kamps, M. Lehtonen, R. Schenkel, J. A. Thom, and A. Trotman. Overview of the INEX 2009 ad hoc track. In Geva et al. [4].

[4] S. Geva, J. Kamps, and A. Trotman, editors. *Focused Retrieval and Evaluation : 8th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2009)*, LNCS. Springer Verlag, Berlin, Heidelberg, 2010.

[5] W.-C. Huang, S. Geva, and A. Trotman. Overview of the INEX 2009 link the wiki track. In Geva et al. [4].

[6] N. Jardine and C. J. van Rijsbergen. The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

[7] G. Kazai, A. Doucet, M. Koolen, and M. Landoni. Overview of the INEX 2009 book track. In Geva et al. [4].

[8] G. Kazai, N. Milic-Frayling, and J. Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32nd Annual International ACM SIGIR Conference*. ACM Press, 2009.

[9] A. Louis and A. Nenkova. Performance confidence estimation for automatic summarization. In *EACL*, pages 541–548. The Association for Computer Linguistics, 2009.

[10] V. Moriceau, E. SanJuan, X. Tannier, and P. Bellot. Overview of the INEX 2009 question answering track: A common task for QA, focused IR and automatic summarization systems. In Geva et al. [4].

[11] R. Nayak. XML data mining: Process and applications. In M. Song and Y.-F. Wu, editors, *Handbook of Research on Text and Web Mining Technologies*. Idea Group Inc., USA, 2008.

[12] R. Nayak, C. M. De Vries, S. Kutty, S. Geva, L. Denoyer, and P. Gallinari. Overview of the INEX 2009 XML mining track: Clustering and classification of XML documents. In Geva et al. [4].

[13] N. Pharo, R. Nordlie, N. Fuhr, T. Beckers, and K. N. Fachry. Overview of the INEX 2009 interactive track. In Geva et al. [4].

[14] 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.

[15] R. Schenkel and M. Theobald. Overview of the INEX 2009 efficiency track. In Geva et al. [4].

[16] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA, 2007. ACM Press.

[17] M. Theobald and R. Schenkel. Overview of the inex 2008 efficiency track. In S. Geva, J. Kamps, and A. Trotman, editors, *7th INEX Workshop*, volume 5631 of *Lecture Notes in Computer Science*, pages 179–191. Springer, 2009.

[18] E. Voorhees. The TREC question answering track. *Journal of Natural Language Engineering*, 7:361–378, 2001.

[19] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference*, pages 603–610. ACM, 2008.