Report on INEX 2011

P. Bellot T. Chappell A. Doucet S. Geva J. Kamps
G. Kazai M. Koolen M. Landoni M. Marx V. Moriceau
J. Mothe G. Ramírez M. Sanderson E. Sanjuan F. Scholer
X. Tannier M. Theobald M. Trappett A. Trotman Q. Wang

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 2011 evaluation campaign, which consisted of a five active tracks: Books and Social Search, Data Centric, Question Answering, Relevance Feedback, and Snippet Retrieval. INEX 2011 saw a range of new tasks and tracks, such as Social Book Search, Faceted Search, Snippet Retrieval, and Tweet Contextualization.

1 Introduction

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

INEX 2011 was an exciting year for INEX in which a number of new tasks and tracks started, including Social Search, Faceted Search, Snippet Retrieval, and Tweet Contextualization. In total five research tracks were included, which studied different aspects of focused information access:

Books and Social Search Track investigating techniques to support users in searching and navigating books, metadata and complementary social media. The *Social Search for Best Books Task* studies the relative value of authoritative metadata and user-generated content using a collection based on data from Amazon and LibraryThing. The *Prove It Task* asks for pages confirming or refuting a factual statement, using a corpus of the full texts of 50k digitized books.

Data Centric Track investigating retrieval over a strongly structured collection of documents based on IMDb. The *Ad Hoc Search Task* has informational requests to be answered by the entities in IMDb (movies, actors, directors, etc.). The *Faceted Search Task* asks for a restricted list of facets and facet-values that will optimally guide the searcher toward relevant information.

Question Answering Track investigating tweet contextualization, answering questions of the form "what is this tweet about?" with a synthetic summary of contextual information grasped from Wikipedia and evaluated by both the relevant text retrieved, and the "last point of interest."

Relevance Feedback Track investigating the utility of incremental passage level relevance feedback by simulating a searcher's interaction. An unconventional evaluation track where submissions are executable computer programs rather than search results.

Snippet Retrieval Track investigating how to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself.

In the rest of this paper, we discuss the aims and results of the INEX 2011 tracks in relatively self-contained sections: the Books and Social Search track (Section 2), the Data Centric track (Section 3), the Question Answering track (Section 4), the Relevance Feedback track (Section 5), and the Snippet Retrieval track (Section 6).

2 Books and Social Search Track

In this section, we will briefly discuss the aims, results and future plans of the INEX 2011 Books and Social Search Track. Further details are in [5].

2.1 Aims and Tasks

The goal of the INEX Book Track is to evaluate techniques for supporting users in searching, navigating and reading book metadata and full texts of digitised books. Toward this goal, the track provides opportunities to explore research questions around four areas: The *Social Search for Best Books* (SB) task, framed within the user task of searching a large online book catalogue for a given topic of interest, aims at comparing retrieval effectiveness from traditional book descriptions, e.g., library catalogue information, and user-generated content such as reviews, ratings and tags. The *Prove It* (PI) task aims to test focused retrieval approaches on collections of books, where users expect to be pointed directly at relevant book parts that may help to confirm or refute a factual claim. The *Structure Extraction* (SE) task aims at evaluating automatic techniques for deriving structure from OCR data and building hyperlinked table of contents. The *Active Reading* (ART) task aims to explore suitable user interfaces to read, annotate, review, and summarise multiple books.

A total of 47 organizations registered for the track, but only 10 groups took an active role. In this section, we will briefly discuss the SB and PI tasks. Further details on all four tasks are available in [5].

2.2 Test Collections

The SB task was newly introduced this year and is centered around the Amazon+LibraryThing collection of 2.8 million book descriptions in XML, containing library catalogue information as well as user-generated content (UGC) in the form of ratings, reviews and tags from Amazon and LibraryThing. Searchers are no longer limited to the information in library catalogues, but can use descriptions and opinions of other readers to search for and select books. Currently, few search systems allow to search on UGC, and the aim of the SB task is to investigate its value for book search. The LibraryThing discussion forums are used as a source of book requests and 211 discussion topics across a broad range of subject and genres were selected for the test collection. We use the recommendation from other members as relevance judgments for evaluation. The topics from LibraryThing reveal an interesting aspect of book search in a social environment. Readers ask for recommendations that are readable for particular age groups, have good story lines or fun characters, books that are engaging or though-provoking. Such aspects are not (well) covered by traditional metadata.

The PI task builds on a collection of over 50,000 digitized out-of-copyright books of different genre (e.g., history books, text books, reference works, novels and poetry) marked up in XML. The task was first run in 2010 and was kept the same for 2011, including the 21 topics for evaluation. Most topics consist of complex factual statements with multiple atomic facts. This creates difficulties for judging, as individual parts can be confirmed or refuted on a page without the other parts being mentioned. In addition to the 6,537 pages judged last year, this year 535 pages were judged by one participating team and 419 pages were judged using AMT, to ensure all top 10 results of all official submissions were judged.

2.3 Results

Four teams together submitted 22 runs for the SB task. We selected 12 fiction-related and 12 non-fiction-related topics for further analysis using AMT. The top 10 results of the 22 runs were pooled and we ask workers on Amazon Mechanical Turk (AMT) to judge books both on topical relevance and whether they are good enough to recommend as the best books on the topic. The official evaluation measure is nDCG@10.

Results show there is not much difference in performance on fiction and non-fiction book requests. However, the relevance judgment from AMT lead to a different system ranking than the suggestions from the LT forums. This supports the intuition that the requests from the LT forums represent a different scenario than ad hoc search for topical relevance. Preliminary analysis also suggests that workers treated relevance and recommendation similarly, although for recommendation the Amazon reviews are much more effective than other parts of the book descriptions. Perhaps the similarity is due to the setup of the experiment, where we asked a single worker to judge both relevance and recommendation, encouraging workers to make judgments dependent upon each other.

Two teams together submitted 21 runs for the PI task. The team from University of Massachusetts submitted the best runs, beating both participants from this and last year. Their best run always returned a confirming or refuting page at the first rank and on average of 6 confirm/refute pages in the top 10.

2.4 Outlook

In this first year of the Social Search for Best Books task we found little difference between fiction and non-fiction book requests and between relevance and recommendation. Next year we will ask assessors to make judgments on specific aspects of relevance independently. We want to explicitly look at aspects such as reading level, interestingness, genre, topical relevance and recommendation. For the Prove It task we will make the topics more structured, such that atomic parts of complex statements can be judged individually.

3 Data Centric Track

In this section, we will briefly discuss the INEX 2011 Data Centric Track. Further details are in [12].

3.1 Aims and Tasks

The goal of the Data Centric Track is to investigate retrieval techniques on highly structured XML data, where the structure is rich and carries semantic information about objects and their relationships. With richly structured XML data, we may ask how well such structural information could be exploited to improve ad hoc retrieval performance and how it could be used in helping users navigate or explore large sets of results as in a faceted search system.

The INEX 2011 Data-Centric Track uses the same IMDB data collection as in 2010, which was generated from the IMDB dump on April 10, 2010. It features two tasks: the ad hoc search task and the faceted search task. The ad hoc search task consists of informational requests to be answered by the entities in the IMDB collection (movies, actors, directors, etc.); the faceted search task asks for a restricted hierarchy of facet-values that will optimally guide the searcher towards relevant information, which is especially useful when searchers information needs are vague or complex.

3.2 Test Collection

Each participating group was asked to create a set of candidate topics, representative of real user needs. We asked participants to submit challenging topics, i.e., topics that could not be easily solved by a current search engine or DB system. Both Content Only (CO) and Content And Structure (CAS) variants of the information need were requested. As for the faceted search task, each topic consists of a general topic as well as a subtopic that refines the general topic by specifying a particular interest of it. For example, animation is a general topic and animation fairy-tale could be a subtopic for it. In total, 8 participating groups submitted 25 valid ad hoc topics and 15 valid general topics along with their 20 subtopics. We added the 20 subtopics to the ad hoc task. Thus, altogether we got 45 topics for the ad hoc search task and 15 topics for the faceted search task. Among them, 38 ad hoc topics were assessed by the participants who submitted runs. The relevance result of a general topic was thought to be that of all or any of its subtopics. So a total of 38 ad hoc topics and 13 faceted search topics were used in evaluations.

Each ad hoc run contains a maximum of 1,000 results per topic in the TREC format. Each faceted run consists of two files: one is the result file containing a ranked list of maximum 2,000 results per topic in the TREC format, and the other is the recommended facet-value file, which can be a static hierarchy of facet-values in XML format or a dynamic faceted search module. A reference result file generated using XPath and Lucene was provided by the organizers so that participants could submit only a facet-value file based on the reference result file.

3.3 Results

A total of 9 active participants submitted 35 ad hoc search runs and 13 faceted search runs. The TREC MAP metric, as well as P@5, P@10, P@20 and so on, was used to measure the performance of all ad hoc runs as whole document retrieval. The best performing runs were submitted by University of Amsterdam and Renmin University of China. They both used the traditional language modeling approaches with no structural information taken into account. However the runs that used structured retrieval models seemed promising in terms of early precisions.

For the faceted search task, since it is the first year, we used two metrics to evaluate the effectiveness of recommended facet-values to gain better understanding to the problem. One metric is the NDCG of facet-values [12]. The NDCG of facet-values is calculated as the NDCG of the results covered by these facet-values. Since in computing the NDCG we considered only the top ten results covered by each facet-value and also limited the number of facet-values to be evaluated to 10, most runs were measured as 0. This could be alleviated by computing the NDCG of the whole hierarchy. The other metric is the interaction cost based on a simple user simulation model [10]. We define users interaction cost as the number of results, facets or facet-values that the user examined before he/she encounters the first relevant result, which is similar to the Reciprocal Rank metric in traditional IR. The best run measured by the interaction cost was from University of Amsterdam, which was based on a result file generated by Indri however. Of all the runs based on the reference result file, the best one was from Renmin University of China, which used a simple redundancy-based approach to recommend facet-values.

3.4 Outlook

The track will be run in 2012, with the name changed to Linked Data Track since we switch the data collection from IMDB to Wikipedia pages plus DBpedia data. We expect to see more complex tasks that require closer interconnection between IR and DB techniques.

4 Question Answering Track

In this section, we will briefly discuss the INEX 2011 Question Answering Track that focused on contextualizing tweets. Further details are in [8].

4.1 Aims and Tasks

Since 2008, Question Answering (QA) track at INEX moved into an attempt to bring together Focused Information Retrieval (FIR) intensively experimented in other INEX tracks (previous ad-hoc tracks and this year snippet track) on the one hand, and topic oriented summarization tasks as defined in NIST Text Analysis Conferences (TAC) [2] on the other hand. Like in recent FIR INEX tasks, the corpus is a clean XML extraction of the content of a dump from Wikipedia. However QA track at INEX differs from current FIR and TAC summarization tasks on the evaluation metrics they use to measure both informativeness and readability. Following [6, 7], informativeness measure is based on lexical overlap between a pool of relevant passages (RPs) and participant summaries. Once the pool of relevant passages is constituted, the process is automatic and can be applied to unofficial runs. The release of these pools is one of the main contributions of INEX QA track. By contrast, readability evaluation is completely manual and cannot be reproduced on unofficial runs. It is based on questionnaires pointing out possible syntax problems, broken anaphora, massive redundancy or other major readability problems.

Therefore QA tasks at INEX moved from the usual IR *query / document* paradigm towards *information need / text answer*. More specifically, the task to be performed by the participating groups of INEX 2011 was contextualizing tweets, *i.e.*, answering questions of the form "what is this tweet about?". Answers could contain up to 500 words and needed to be a concatenation of textual passages from the Wikipedia dump.

4.2 Results

This year, participants had to contextualize 132 tweets from the New York Times (NYT). Informativeness of answers has been evaluated, as well as their readability. 13 teams from 6 countries actively participated to this track.

Informativeness evaluation has been performed by organizers on a pool of 50 topics. For each of these topics, all passages submitted have been evaluated. To check that the resulting pool of relevant answers is sound, a second automatic evaluation for informativeness of summaries has been carried out with respect to a reference made of the NYT article corresponding to the topic.

Like in the 2010 edition, we intended to use the Kullback-Leibler divergence as described in [6, 7] to evaluate the informativeness of short summaries based on a bunch of highly relevant documents. However, in 2010, references were made of complete Wikipedia pages, therefore the textual content was much longer than summaries and smoothing did not introduce too much noise. This is not the case with the 2011 assessments. For some topics, the amount of relevant passages is very low, less than the maximal summary length. We thus simply considered absolute log-diff between frequencies. Dissimilarity values are very closed, however differences are often statistically significant.

A complete baseline system based on a Indri Index and TreeTagger was provided to participants. The three top ranked runs did not use it meanwhile all other runs improving the baseline used it only to query the Indri Index, some applying special query expansion techniques. None of the participants used this year the baseline summarization system which ranks 7th among all runs when returning full sentences and 19th when returning only noun phrases.

4.3 Outlook

The tweet contextualization task will continue at INEX 2012 with same methodology and baseline but on a much wider range of tweet types.

5 Relevance Feedback Track

In this section, we will briefly discuss the INEX 2011 Relevance Feedback track. Further details are in [1].

5.1 Aims and Task

The Relevance Feedback track, run for the second time, involves the submission of relevance feedback algorithms by participants. These algorithms interface with an evaluation platform made available to participating organizations which simulates an end user of a search engine.

The relevance feedback algorithms are presented with a number of topics and for each topic they must provide the user with the next most relevant document that has not already been returned. After seeing each document, the evaluation platform will inform the algorithm which parts of the document, if any, are relevant to the search. The algorithm may then use this information to rerank results and provide more relevant documents in the future.

5.2 Test Collection

This track was run in identical fashion to the INEX 2010 Relevance Feedback track. Participating organizations submit a *relevance feedback module* (RFM) in the form of a Java library. Each RFM implements a particular algorithm for ranking and reranking a list of provided documents according to feedback passed to it. The RFMs are each evaluated by an *evaluation platform* (EP) that links the RFMs, provides them with a hitherto unknown set of documents and simulates a user making use of the system, looking through the documents returned by the RFM and providing relevance judgments based on preexisting relevance assessments. The RFMs are then scored based on how well they returned relevant information to the simulated user.

The main innovation between the INEX 2010 and 2011 versions of the track was the availability of a native client module that would interface with a relevance feedback module written in another programming language. This was to address the issues involved with requiring all relevance feedback modules to be written in Java.

5.3 Results

Two groups submitted a total of four Relevance Feedback Modules to the 2011 version of the track (down from 10 groups in 2010) due to the track running particularly late. QUT resubmitted the reference Relevance Feedback Module described in the next paragraph while the University of Otago submitted three native client submissions using the supplied driver. As with the INEX 2010 version of the track, a Lucene-based reference module was provided by QUT. This reference module used a scrolling character buffer which was filled with relevant passages and filtered for popular terms (ranked by term frequency), which were used to expand the search query. The University of Otago made three submissions of a native client that uses the ATIRE search engine with various settings, including Rocchio pseudo-relevance feedback and tuning. These submissions did not incorporate the supplied feedback, leaving the reference module as the only submission to use this information. The basic module, which incorporated BM25, and the Rocchio pseudo-relevance feedback module achieved nearly identical performance while the tuned version achieved greater early precision.

5.4 Outlook

With the availability of a wide range of effective IR platforms (Lucene, Indry, Terrier, Zettair, Wumpus, Ant, etc.) fewer IR researchers are still building their own systems from scratch. The Feedback Track will continue at INEX 2012, but with the more general aim to promote system and component building by participants, by giving a platform to “show your code.”

6 Snippet Retrieval Track

In this section, we will briefly discuss the INEX 2011 Snippet Retrieval Track. Further details are in [11].

6.1 Aims and Task

The goal of the snippet retrieval track has been to determine how best to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself, allowing the user to quickly find what they are looking for. The task was to return a ranked list of documents for the requested topic to the user, and with each document, a corresponding text snippet describing the document. Each run was allowed to return up to 500 documents per topic, with a maximum of 300 characters per snippet.

6.2 Test Collection

The Snippet Retrieval Track uses the INEX Wikipedia collection introduced in 2009 — an XML version of the English Wikipedia, based on a dump taken on 8 October 2008, and semantically annotated as described by Schenkel et al. [9]. This corpus contains 2,666,190 documents. A set of 50 topics were taken from the INEX 2009 Ad Hoc Track [3]. Each topic contains a short content only (CO) query, a content and structure (CAS) query, a phrase title, a one line description of the search request, and a narrative with a detailed explanation of the information need, the context and motivation of the information need, and a description of what makes a document relevant or not.

To determine the effectiveness of the returned snippets at their goal of allowing a user to determine the relevance of the underlying document, manual assessment has been used. The documents for each topic were manually assessed for relevance based on the snippets alone, as the goal is to determine the snippet's ability to provide sufficient information about the document. Each topic within a submission was assigned an assessor. The assessor, after reading the details of the topic, read through the top 100 returned snippets, and judged which of the underlying documents seemed relevant based on the snippets. To avoid bias introduced by assessing the same topic more than once in a short period of time, and to ensure that each submission is assessed by the same assessors, the runs were shuffled in such a way that each assessment package contained one run from each topic, and one topic from each submission.

Submissions were evaluated by comparing the snippet-based relevance judgements with the existing document-based relevance judgements, which were treated as a ground truth. The primary evaluation metric used is the geometric mean of recall and negative recall (GM). A high value of GM requires a high value in recall and negative recall — i.e. the snippets must help the user to accurately predict both relevant and irrelevant documents. If a submission has high recall but zero negative recall (e.g. in the case that everything is judged relevant), GM will be zero. Likewise, if a submission has high negative recall but zero recall (e.g. in the case that everything is judged irrelevant), GM will be zero.

6.3 Results

In total, 44 runs were accepted from a total of 56 runs submitted. These runs came from 9 different groups, based in 7 different countries. The highest ranked run was ‘p72-LDKE-1111’, submitted by the Jiangxi University of Finance and Economics. No run scored higher than 47% in recall, with an average of 35%. This indicates that poor snippets were causing users to miss more than half of all relevant results. Negative recall was high, with no run scoring below 80%, signifying that users are able to identify most irrelevant results based on snippets. We refer to [11] for further details and analysis of the results.

6.4 Outlook

The captioning problem in IR is far from solved, and the snippet retrieval track will continue at INEX 2012.

7 Envoi

This complete our walk-through of the five tracks of INEX 2011. 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 2011 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 2011, however, is a great number of test collections that can be used for future experiments.

INEX 2012 will see some exciting changes. After ten INEX workshops organized as a stand-alone event, INEX decided to team up with CLEF, and the INEX 2012 workshop will be held during CLEF in Rome on September 17–20. The schedule for INEX 2012 is therefore earlier than in previous years, and at the time of writing INEX 2012 has started and will continue the tracks of INEX 2011 with a number of exciting innovations. The Social Book Search track will continue exploring the relative value of formal description and social media for search and recommendation. The Data Centric Track will evolve into a Linked Data Track focusing on `http://dbpedia.org/` and `http://en.wikipedia.org/`. The Question Answering track lost it’s or original QA focus and continue with tweet or post contextualization. The Feedback track will continue and widen its scope to other IR system components—e.g., “show me your code.” And the Snippet track will continue to look at effective forms of captioning conveying the relevance of complex search results.

References

[1] T. Chappell and S. Geva. Overview of the INEX 2011 relevance feedback track. In Geva et al. [4].

[2] H. Dang. Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In *Proc. of the First Text Analysis Conference*, 2008.

[3] S. Geva, J. Kamps, M. Lethonen, R. Schenkel, J. Thom, and A. Trotman. Overview of the INEX 2009 ad hoc track. In S. Geva, J. Kamps, and A. Trotman, editors, *Focused Retrieval and Evaluation*, volume 6203 of *LNCS*, pages 4–25. Springer, 2010.

[4] S. Geva, J. Kamps, and R. Schenkel, editors. *Focused Retrieval of Content and Structure: 10th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2011)*, LNCS. Springer, 2012.

[5] M. Koolen, G. Kazai, J. Kamps, A. Doucet, and M. Landoni. Overview of the INEX 2011 books and social search track. In Geva et al. [4].

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

[7] H. Saggion, J.-M. Torres-Moreno, I. da Cunha, E. SanJuan, and P. Velázquez-Morales. Multilingual summarization evaluation without human models. In C.-R. Huang and D. Jurafsky, editors, *COLING (Posters)*, pages 1059–1067. Chinese Information Processing Society of China, 2010.

[8] E. Sanjuan, V. Moriceau, X. Tannier, P. Bellot, and J. Mothe. Overview of the INEX 2011 question answering track. In Geva et al. [4].

[9] 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.

[10] A. Schuth and M. Marx. Evaluation methods for rankings of facetvalues for faceted search. In *Multilingual and Multimodal Information Access Evaluation*, volume 6941 of *LNCS*, pages 131–136, 2011.

[11] M. Trappett, S. Geva, A. Trotman, M. Sanderson, and F. Scholer. Overview of the INEX 2011 snippet retrieval track. In Geva et al. [4].

[12] Q. Wang, G. Ramírez, M. Marx, M. Theobald, and J. Kamps. Overview of the INEX 2011 data centric track. In Geva et al. [4].