Overview of INEX 2014

Patrice Bellot, Toine Bogers, Shlomo Geva, Mark Hall, Hugo Huurdeman, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Véronique Moriceau, Josiane Mothe, et al.

Overview of INEX 2014

Patrice Bellot, Toine Bogers, Shlomo Geva[1], Mark Hall, Hugo Huurdeman, Jaap Kamps[2], Gabriella Kazai, Marijn Koolen, Véronique Moriceau, Josiane Mothe, Michael Preminger, Eric SanJuan, Ralf Schenkel[3], Mette Skov, Xavier Tannier, and David Walsh

[1] INEX co-chair & QUT, Australia
[2] INEX co-chair & University of Amsterdam, The Netherlands
[3] INEX co-chair & University of Passau, Germany

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 2014 evaluation campaign, which consisted of three tracks: The *Interactive Social Book Search Track* investigated user information seeking behavior when interacting with various sources of information, for realistic task scenarios, and how the user interface impacts search and the search experience. The *Social Book Search Track* investigated the relative value of authoritative metadata and user-generated content for search and recommendation using a test collection with data from Amazon and LibraryThing, and user profiles and personal catalogues. The *Tweet Contextualization Track* investigated tweet contextualization, helping a user to understand a tweet by providing him with a short background summary generated from relevant Wikipedia passages aggregated into a coherent summary. INEX 2014 was an exciting year for INEX in which we for the third time ran our workshop as part of the CLEF labs in order to facilitate knowledge transfer between the evaluation forums. This paper gives an overview of all the INEX 2014 tracks, their aims and task, the built test-collections, the participants, and gives an initial analysis of the results.

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 2014 was an exciting year for INEX in which we further integrated into the CLEF Labs structure in order to foster further collaboration and facilitate knowledge transfer between the evaluation forums. In total three research tracks were included, which studied different aspects of focused information access:

Interactive Social Book Search Track investigates user information seeking behavior when interacting with various sources of information, for realistic task scenarios, and how the user interface impacts search and the search experience.

Social Book Search Track investigates the relative value of authoritative metadata and user-generated content for search and recommendation using a test collection with data from Amazon and LibraryThing, and user profiles and personal catalogues.

Tweet Contextualization Track investigates tweet contextualization, helping a user to understand a tweet by providing him with a short background summary generated from relevant Wikipedia passages aggregated into a coherent summary (in collaboration with the RepLab Lab).

Also a continuation of the Linked Data Track was announced (in collaboration with the CLEF QA Lab), in particular the Jeopardy Task running SPARQL queries on a DBpedia/Wikipedia corpus, but eventually the QALD task opted for a different corpus.

In the rest of this paper, we discuss the aims and results of the INEX 2014 tracks in relatively self-contained sections: the Interactive Social Book Search track (Section 2), the Social Books Search track (Section 3), and the Tweet Contextualization (Section 4) track.

2 Interactive Social Book Search Track

In this section, we will briefly discuss the INEX 2014 Interactive Social Book Search Track. Further details are in [4].

2.1 Aims and Tasks

The goal of the Interactive Social Book Search (ISBS) track is to investigate how book searchers use professional metadata and user-generated content at different stages of the search process. The purpose of this task is to gauge user interaction and user experience in social book search by observing user activity with a large collection of rich book descriptions under controlled and simulated conditions, aiming for as much "real-life" experiences intruding into the experimentation. The output will be a rich data set that includes both user profiles, selected individual differences (such as a motivation to explore), a log of user interactivity, and a structured set of questions about the experience.

The Interactive Social Book Search Track is a merger of the INEX Social Book Search Track (discussed in Section 3 below) and the Interactive task of CHiC [7, 9]. The SBS Track started in 2011 and has focused on system-oriented evaluation of book search systems that use both professional metadata and user-generated content. Out of three years of SBS evaluation arose a need to understand how users interact with these different types of book descriptions and how systems could support user to express and adapt their information needs during

the search process. The CHiC Interactive task focused on interaction of users browsing and searching in the Europeana collection. One of the questions is what types of metadata searchers use to determine relevance and interest. The collection, use case and task were deemed not interesting and useful enough to users. The first year of the ISBS will focus on switching to the SBS collection and use case, with as few other changes as possible.

The goal of the interactive book search task is to investigate how searchers interact with book search systems that offer different types of book metadata. The addition of opinionated descriptions and user-supplied tags allows users to search and select books with new criteria. User reviews may reveal information about plot, themes, characters, writing style, text density, comprehensiveness and other aspects that are not described by professional metadata. In particular, the focus is on complex goal-oriented tasks as well as non-goal oriented tasks. For traditional tasks such as known-item search, there are effective search systems based on access points via formal metadata (i.e. book title, author name, publisher, year, etc). But even here user reviews and tags may prove to have an important role. The long-term goal of the task is investigate user behavior through a range of user tasks and interfaces and to identify the role of different types of metadata for different stages in the book search process.

For the Interactive task, the main research question is: *How do searchers use professional metadata and user-generated content in book search?* This can be broken down into a few more specific questions:

RQ1 *How should the system and user interface combine professional and user-generated information?*

RQ2 *How should the system adapt itself as the user progresses through their search task?*

2.2 Experimental Setup

The track builds on the INEX Amazon/LibraryThing (A/LT) collection [1, see also Section 3], which contains 1.5 million book descriptions from Amazon, enriched with content from LT. This collection contains both professional metadata and user-generated content. This collection is a subset of a larger collection of 2.8 million description, selecting all and only book descriptions that have a cover image.

Two tasks were created to investigate the impact of different task types on the participants interactions with the interfaces and also the professional and user-generated book meta-data. The first is a *goal-oriented* task, developed as a "simulated leisure task" [8] based on a topic derived from the LibraryThing discussion fora:

> Imagine you are looking for some interesting physics and mathematics books for a layperson. You have heard about the Feynman books but you have never really read anything in this area. You would also like to find an "interesting facts" sort of book on mathematics.

The LibraryThing collection contains discussion fora in which users asked other users for advice on which books to read for a given topic, question, or area of interest. From this list of discussion topics, a discussion on "layman books for physics and mathematics" was selected as the book collection contained a significant number of books on the topic, it is a neutral topic, it provides guidance, but it is also sufficiently flexible that participants can interpret it as needed.

The second is a *non-goal-oriented* task, based on the open-ended task used in the iCHiC task at CLEF 2013 [9]:

> Imagine you are waiting to meet a friend in a coffee shop or pub or the airport or your office. While waiting, you come across this website and explore it looking for any book that you find interesting, or engaging or relevant...

The aim of this task is to investigate how users interact with the system when they have no pre-defined goal in a more exploratory search context. It also allows the participants to bring their own goals or sub-tasks to the experiment in line with the "simulated work task" ideas [3].

The setup used extensive questionnaires as fascinated by the SPIRE system [9]: *Consent* questionnaire: all participants had to confirm that they understood the tasks they would be asked to undertake and the types of data collected in the experiment, and also specified who had recruited them; *Demographics* questionnaire: the following factors were acquired in order to characterize the participants: gender, age, achieved education level, current education level, and employment status; *Culture* questionnaire: to quantify language and cultural influences, the following factors were collected: country of birth, country of residence, mother tongue, primary language spoken at home, languages used to search the web; *Post-Task* questionnaire: in the post task questions, participants were asked to judge how useful each of the interface components and meta-data parts that they had used in the task were, using 5-point Likert-like scales; and *Engagement* questionnaire: after participants had completed both tasks, they were asked to complete O'Brien and Toms [6]'s engagement scale.

Two distinct systems were developed. The first is a *Baseline* system representing a standard web-search interface, with the left column containing the task instructions, book-bag, and search history and the main area showing the results, see Figure 1.

The second is a *Multistage* system, having different views for three stages of the search process, see Figure 2. The initial *explore* stage aimed to support the initial exploration of the data-set and contains a very similar feature set to the baseline, including task instructions, search box, search results, book bag, and search history. The two main differences to the *baseline* interface were the navigation bar that allows the participants to switch between the stages and the dense, multi-column search results. The *focus* stage supports in-depth searching and provides detailed search results that directly include the full meta-data that in the other stages is shown via a popup. A category filter was also provided in the left column which provided a means to reduce and refine the search results. The *refine* stage supports the refining of the final list of books the participants

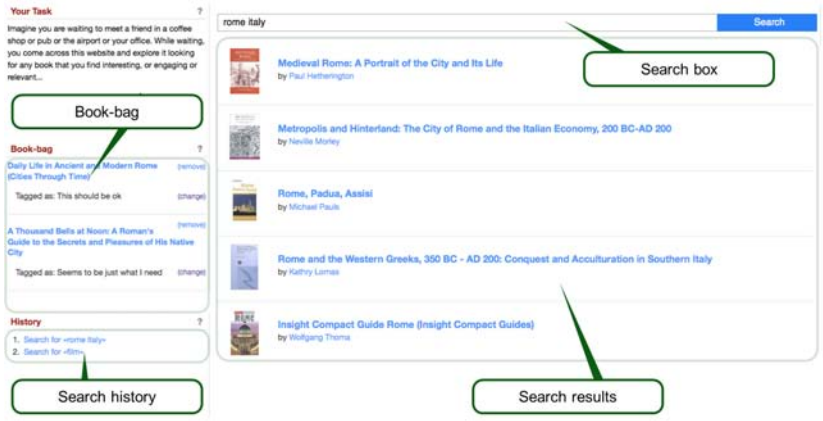

Fig. 1. Baseline interface's results view

Table 1. Overview of the participating teams and number of users per team

Institute	# Test persons
Aalborg	7
Amsterdam	7
Edge Hill	10
Humboldt	17
Total	41

want to choose. It thus focuses on the books the user has already added to their book-bag and this stage cannot be entered until at least one book has been added to the book-bag.

2.3 Results

A total of four teams contributed 41 test persons to the experiments. In Table 1 we show which institutes participated in this track and the number of users that took part in their experiments.

Based on the participant responses and log data we have aggregated summary statistics for a number of basic performance metrics in Table 2.

Session length shows median and inter-quartile ranges in minutes and seconds for all interface and task combinations. While the results seem to indicate that participants spent longer in the *Baseline* interface and also longer on the *goal-oriented* task, the differences are not statistically significant (Wilcoxon signed-rank test). For the *non-goal* task, the median times are roughly similar to the session lengths in the iCHiC experiments This might indicate that that is the approximate time that participants can be expected to spend on any kind of open-ended leisure-task.

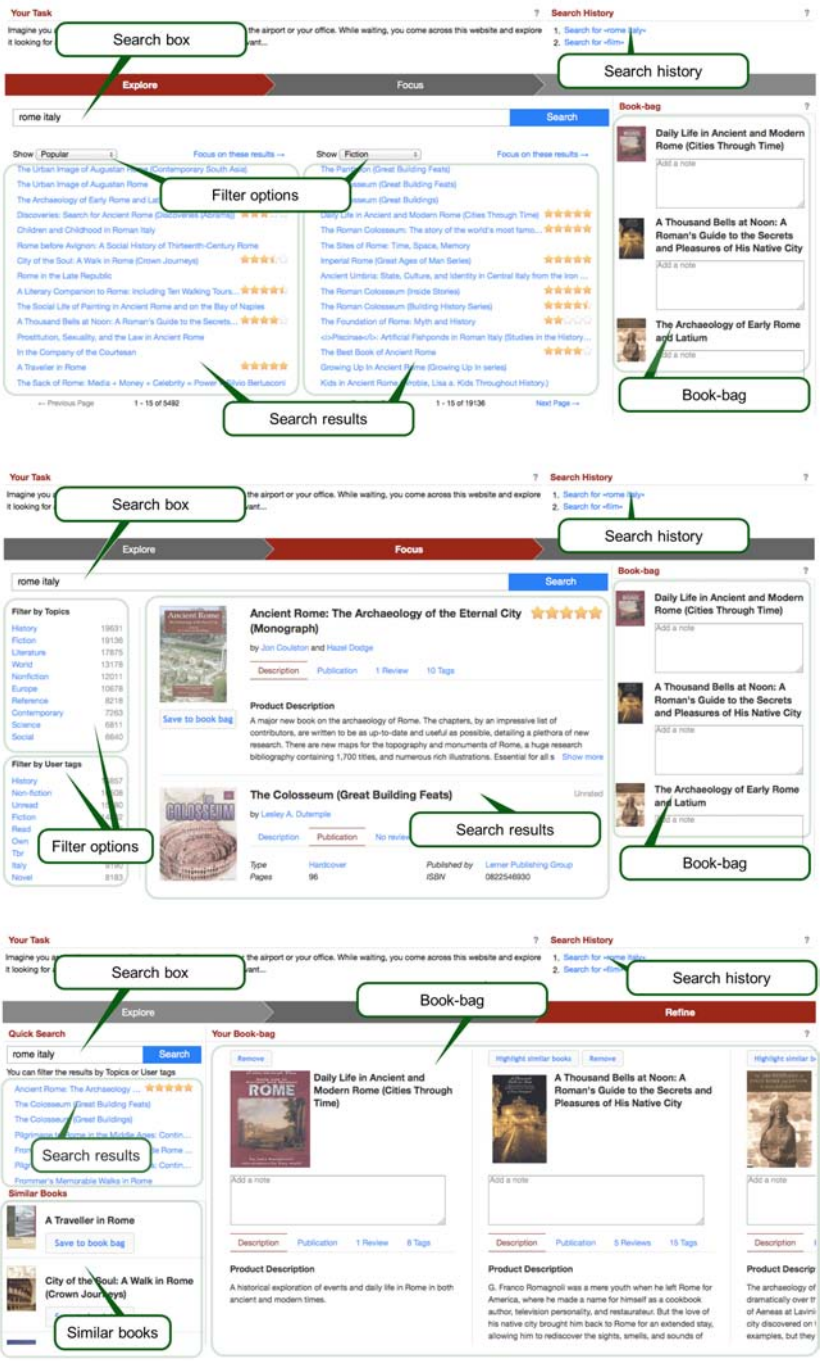

Fig. 2. Multistage interface: Explore view (top), Focus view (middle), and Refine view (bottom)

Table 2. Statistics over systems and tasks

	Goal-oriented		**Non-goal**	
Session Length				
Baseline	6:25min	(3:42min)	3:42min	(3:45min)
Multi-Stage	3:35min	(4:24min)	2:40min	(6:21min)
Number of Queries				
Baseline	4	(5.5)	2	(4.5)
Multi-Stage	3	(2.75)	2	(3)
Number of Books Viewed				
Baseline	4	(5.5)	2	(4.5)
Multi-Stage	3	(2.75)	2	(3)
Number of Books Collected				
Baseline	3	(3)	1	(2)
Multi-Stage	3.5	(3)	2	(3)

Number of queries shows median and inter-quartile ranges for each interface and task. The results are in line with the session length results, with participants executing slightly more queries in the *goal-oriented* task (Wilcoxon rank-sum test $p < 0.05$). However, the interface did not have a significant impact on the number of queries executed.

Number of books viewed shows median and inter-quartile ranges for each interface and task. Participants viewed fewer books in the *non-goal* task (Wilcoxon rank-sum test $p < 0.05$), which was to be expected considering that they also executed less queries and spent less time on the task. As with the number of queries the number of books viewed is not significantly influenced by the interface participants used.

Number of books collected shows median and inter-quartile ranges for each combination, based on the number of books participants had in their book-bag when they completed the session, not the total number of books collected over the course of their session. Participants collected those books that they felt were of use to them. Unlike the other metrics, where the interface had no significant influence on the metric, in the *non-goal* task, participants collected significantly more books using the *multi-stage* interface than with the *baseline* interface. Considering that there are no significant interface effects for the *non-goal* task in any of the other metrics and that there is no significant difference in the *goal-oriented* task, this strongly suggests that the *multi-stage* interface provides a benefit to open-ended leisure tasks, while at the same time working just as well as the *baseline* interface for more focused tasks.

2.4 Outlook

As the focus on the INEX 2013 Interactive Social Book Search track was switching to the SBS collection and use case, in particular in terms of the experimental

systems and the infrastructure to collect log and questionnaire data, the 2013 edition had the character of a pilot track. Next year, we are able to reap the benefits of these investments and continue with the ISBS track to further investigate how books searchers use professional metadata and user-generated content at different stages of the search process.

3 Social Book Search Track

In this section, we will briefly discuss the INEX 2014 Social Book Search Track. Further details are in [5].

3.1 Aims and Tasks

For centuries books were the dominant source of information, but how we acquire, share, and publish information is changing in fundamental ways due to the Web. The goal of the Social Book Search Track is to investigate techniques to support users in searching and navigating the full texts of digitized books and complementary social media as well as providing a forum for the exchange of research ideas and contributions. Towards this goal the track is building appropriate evaluation benchmarks, complete with test collections for social, semantic and focused search tasks. The track provides opportunities to explore research questions around two key areas: First, evaluation methodologies for book search tasks that combine aspects of retrieval and recommendation. Second, information retrieval techniques for dealing with professional and user-generated metadata.

The *Social Book Search* (SBS) task, framed within the scenario of a user searching a large online book catalogue for a given topic of interest, aims at exploring techniques to deal with complex information needs—that go beyond topical relevance and can include aspects such as genre, recency, engagement, interestingness, and quality of writing—and complex information sources that include user profiles, personal catalogues, and book descriptions containing both professional metadata and user-generated content.

The 2014 edition represents the fourth consecutive year the SBS task has run and once more the test collection used is the Amazon/LibraryThing collection of 2.8 million documents. LibraryThing forum requests for book suggestions, combined with annotation of these requests resulted in a topic set of 680 topics with graded relevance judgments. Compared to 2013, there are three important changes: (1) a much larger set of 94,000+ user profiles was provided to the participants this year; (2) an additional 300 forum topics were annotated, bringing the total number of topics up to 680; and (3) the *Prove It* task did not run this year. Prompted by the availability of large collections of digitized books, the Social Book Search Track aims to promote research into techniques for supporting users in searching, navigating and reading full texts of digitized books and associated metadata.

3.2 Test Collections

For the Social Book Search task a new type of test collection has been developed. Unlike traditional collections of topics and topical relevance judgements, the task is based on rich, real-world information needs from the LibraryThing (LT) discussion forums and user profiles. The collection consists of 2.8 million book descriptions from Amazon, including user reviews, and is enriched with user-generated content from LT. This collection was originally constructed by Beckers et al. [1], but extended and augmented in various ways, see [5].

For the information needs we used the LT discussion forums. Over the past two years, we had a group of eight different Information Science students annotate the narratives of a random sample of 2,646 LT forum topics. Of the 2,646 topics annotated by the students, 944 topics (36%) were identified as containing a book search information need. Because we want to investigate the value of recommendations, we use only topics where the topic creators add books to their catalogue both before (pre-catalogued) and after starting the topic (post-catalogued). Without the former, recommender systems have no profile to work with and without the latter the recommendation part cannot be evaluated. This leaves 680 topics. These topics were combined with all the pre-catalogued books of the topic creators' profiles and distributed to participating groups. An example of an annotated topic (topic 99309) is:

```
<topic id="99309">
  <query>Politics of Multiculturalism</query>
  <title>Politics of Multiculturalism Recommendations?</title>
  <group>Political Philosophy</group>
  <member>steve.clason</member>
  <narrative> I'm new, and would appreciate any recommended reading on
    the politics of multiculturalism. <a href="/author/parekh">Parekh
    </a>'s <a href="/work/164382">Rethinking Multiculturalism: Cultural
    Diversity and Political Theory</a> (which I just finished) in the end
    left me unconvinced, though I did find much of value I thought he
    depended way too much on being able to talk out the details later. It
    may be that I found his writing style really irritating so adopted a
    defiant skepticism, but still... Anyway, I've read
    <a href="/author/sen">Sen</a>, <a href="/author/rawles">Rawls</a>,
    <a href="/author/habermas">Habermas</a>, and
    <a href="/author/nussbaum">Nussbaum</a>, still don't feel like I've
    wrapped my little brain around the issue very well and would
    appreciate any suggestions for further anyone might offer.
  </narrative>
  <catalog>
    <book>
      <LT_id>9036</LT_id>
      <entry_date>2007-09</entry_date>
      <rating>0.0</rating>
      <tags></tags>
    </book>
    <book>
      ...
```

Table 3. User profile statistics of the topic creators and all other users.

Type	N	total	min	max	median	mean	stdev
Topic Creators							
Pre-catalogued	680	399,147	1	5884	239	587	927
Post-catalogued	680	209,289	1	5619	114	308	499
Total catalogue	680	608,436	2	8563	432	895	1202
All users							
Others	93,976	33,503,999	1	41,792	134	357	704
Total	94,656	34,112,435	1	41,792	135	360	710

The relevance judgements come in the form of suggestions from other LT members in the same discussion thread and the additional annotations, translated into a graded relevance scale (see [5] for details).

In addition to information needs of social book search topics, LT also provides the rich user profiles of the topic creators and other LT users, which contain information on which books they have in their personal catalogue on LT, which ratings and tags they assigned to them and a social network of friendship relations, interesting library relations and group memberships. These profiles may provide important signals on the user's topical and genre interests, reading level, which books they already know and which ones they like and don't like. These profiles were scraped from the LT site, anonymised and made available to participants. Basic statistics on the number of books per user profile is given in Table 3. By the time users ask for book recommendations, most of them already have a substantial catalogue (pre-catalogued). The distribution is skewed, as the mean (587) is higher than the median (239). After posting their topics, users tend to add many more books (post-catalogued), but fewer than they have already added. Compared to the other users in our crawl (median of 134 books), the topic creators are the more active users, with larger catalogues (median of 432 books).

3.3 Results

A total of 64 teams registered for the track (compared with 68 in 2013, 55 in 2012 and 47 in 2011). At the time of writing, we counted 8 active groups (compared with 8 in 2013, 5 in 2012 and 10 in 2011) submitting a total of 40 runs, see Table 4.

The official evaluation measure for this task is nDCG@10. It takes graded relevance values into account and is designed for evaluation based on the top retrieved results. In addition, P@10, MAP and MRR scores will also be reported, with the evaluation results shown in Table 5.

The best performing run is *run6.SimQuery1000.rerank_all.L2R_RandomForest* by USTB, which used all topic fields combined against an index containing all available document fields. The run is re-ranked with 12 different re-ranking

Table 4. Active participants of the INEX 2014 Social Book Search Track and number of contributed runs

ID	Institute	Acronym	Runs
4	University of Amsterdam	UvA	4
54	Aalborg University Copenhagen	AAU	3
65	University of Minnesota Duluth	UMD	6
123	LSIS / Aix-Marseille University	SBS	6
180	Chaoyang University of Technology	CYUT	4
232	Indian School of Mines, Dhanbad	ISMD	5
419	Université Jean Monnet	UJM	6
423	University of Science and Technology Beijing	USTB	6
	Total		40

Table 5. Evaluation results for the official submissions (best run per team). Best scores are in bold. Runs marked with ⋆ are manual runs.

Group	Run	nDCG@10	P@10	MRR	MAP	Profiles
USTB	run6.SimQuery1000.rerank_all.-L2R_RandomForest	**0.303**	**0.464**	**0.232**	0.390	No
UJM	326	0.142	0.275	0.107	0.426	No
LSIS	InL2	0.128	0.236	0.101	0.441	No
AAU	run1.all-plus-query.all-doc-fields	0.127	0.239	0.097	**0.444**	No
CYUT	Type2QTGN	0.119	0.246	0.086	0.340	No
UvA	inex14.ti_qu.fb.10.50.5000	0.097	0.179	0.073	0.421	No
UMD	Full_TQG_fb.10.50_0.0000227_50	0.097	0.188	0.069	0.328	Yes
⋆ISMD	354	0.067	0.123	0.049	0.285	No

strategies, which are then combined adaptively using learning-to-rank. The second group is UJM with run *326*, which uses BM25 on the title, mediated query and narrative fields, with the parameters optimised for the narrative field. The third group is **lsis**, with *InL2*. This run is based on the InL2 model, the index is built from all fields in the book xml files. The system uses the mediated query, group and narrative fields as a query.

There are 11 systems that made use of the user profiles, but they are not among the top ranking systems. The best systems combine various topic fields, with parameters trained for optimal performance. This is the first year that systems included learning-to-rank approaches, the best of which clearly outperforms all other systems.

Last year there were many (126 out of 380, or 33%) topics for which none of the systems managed to retrieve any relevant books. This year, there were only 56 of these topics (8%). There are 27 topics where the only books suggested in the thread are already catalogued or read by the topic creator, so all relevance values are zero. The other 39 topics where all systems fail to retrieve relevant books have very few (mostly 1 or 2) suggestions and tend to be very vague

or broad topics where hundreds or thousands of books could be recommended. This drop is probably due to the restriction of selecting only topics of users who catalogue books. Many of the topics on which all systems fail are known-item topics posed by users who have either a private catalogue or who are new users with empty catalogues. These have been removed from this year's topic pool. By selecting topics from only active users, the evaluation moves further away from known-item search.

3.4 Outlook

This was the fourth year of the Social Book Search Track. The track ran only a single tasks: the system-oriented Social Book Search task, which continued its focus on both the relative value of professional and user-generated metadata and the retrieval and recommendation aspects of the LT forum users and their information needs. Next year, we plan to shift the focus of the SBS task to the interactive nature of the topic thread and the suggestions and responses given by the topic starter and other members. We are also thinking of a pilot task in which the system not only has to retrieve relevant and recommendable books, but also to select which part of the book description—e.g. a certain set of reviews or tags—is most useful to show to the user, given her information need.

4 Tweet Contextualization Track

In this section, we will briefly discuss the INEX 2014 Tweet Contextualization Track. Further details are in [2].

4.1 Aims and Tasks

Tweets (or posts in social media) are 140 characters long messages that are rarely self-content. The Tweet Contextualization aims at providing automatically information—a summary that explains the tweet. This requires combining multiple types of processing from information retrieval to multi-document summarization including entity linking. Running since 2010, the task in 2014 was a slight variant of previous ones considering more complex queries from RepLab 2013. Given a tweet and a related entity, systems had to provide some context about the subject of the tweet from the perspective of the entity, in order to help the reader to understand it.

The Tweet Contextualization's task in 2014 is a slight variant of previous ones and it is complementary to CLEF RepLab. Previously, given a tweet, systems had to help the user to understand it by reading a short textual summary. This summary had to be readable on a mobile device without having to scroll too much. In addition, the user should not have to query any system and the system should use a resource freely available. More specifically, the guideline specified the summary should be 500 words long and built from sentences extracted from a dump of Wikipedia. In 2014 a small variant of the task has been explored,

considering more complex queries from RepLab 2013, but using the same corpus. The new use case of the task was the following: given a tweet and a related entity, the system must provide some context about the subject of the tweet from the perspective of the entity, in order to help the reader answering questions of the form "why this tweet concerns the entity? should it be an alert?".

In the remaining we give details about the English language tweets, and refer the reader to the overview paper [2] for the pilot task in Spanish.

4.2 Test Collection

The official document collection for 2014 was the same as in 2013. Between 2011 and 2013 the corpus did change every year but not the user case. In 2014, the same corpus was reused but the user case evolved. Since 2014 TC topics are a selection of tweets from RepLab 2013, it was necessary to use prior WikiPedia dumps. Some participants also used the 2012 corpus raising up the question of the impact of updating the WikiPedia over these tasks.

Let us recall that the document collection has been built based on yearly dumps of the English WikiPedia since November 2011. We released a set of tools to convert a WikiPedia dump into a plain XML corpus for an easy extraction of plain text answers. The same perl programs released for all participants have been used to remove all notes and bibliographic references that are difficult to handle and keep only non empty Wikipedia pages (pages having at least one section).

The resulting automatically generated documents from WikiPedia dump, consist of a title (`title`), an abstract (`a`) and sections (`s`). Each section has a subtitle (`h`). Abstract and sections are made of paragraphs (`p`) and each paragraph can contain entities (`t`) that refer to other Wikipedia pages.

As tweets, 240 topics have been collected from RepLab 2013 corpus. These tweets have been selected in order to make sure that:

- They contained "informative content" (in particular, no purely personal messages);
- The document collections from Wikipedia had related content, so that a contextualization was possible.

In order to avoid that fully manual, or not robust enough systems could achieve the task, all tweets were to be treated by participants, but only a random sample of them was to be considered for evaluation.

These tweets were provided in XML and tabulated format with the following information:

- the category (4 distinct),
- an entity name from the wikipedia (64 distinct)
- a manual topic label (235 distinct).

The entity name was to be used as an entry point into WikiPedia or DBpedia. The context of the generated summaries was expected to be fully related to this entity. On the contrary, the usefulness of topic labels for this automatic task was and remains an open question at this moment because of their variety.

4.3 Evaluation

Tweet contextualization is evaluated on both informativeness and readability. Informativeness aims at measuring how well the summary explains the tweet or how well the summary helps a user to understand the tweet content. On the other hand, readability aims at measuring how clear and easy to understand the summary is.

The *informativeness* measure is based on lexical overlap between a pool of relevant passages (RPs) and participant summaries. Once the pool of RPs is constituted, the process is automatic and can be applied to unofficial runs. This year's topics included more facets and converting them into queries for a Research Engine was less straightforward. As a consequence, it was not possible to rely on a pooling from participant runs because it would have been too sparse and incomplete, and a thorough manual run by organizers based on the reference system that was made available to all participants. Unofficial runs based on this reference run can be reliably evaluated.

By contrast, *readability* is evaluated manually and cannot be reproduced on unofficial runs. In this evaluation the assessor indicates where he misses the point of the answers because of highly incoherent grammatical structures, unsolved anaphora, or redundant passages. Three metrics were used: **Relaxed metric**, counting passages where the T box has not been checked; **Syntax metric**, counting passages where the S box was not checked either (i.e, the passage has no syntactic problems), and the **Structure (or Strict) metric** counting passages where no box was checked at all. In all cases, participant runs were ranked according to the average, normalized number of words in valid passages.

4.4 Results

In the 2014 edition of the track, four combined teams from six countries (Canada, France, Germany, India, Russia, Tunesia) submitted 12 runs to the Tweet Contextualization track. Two other teams from Mexico and Spain participated to the pilot task in Spanish submitting three runs as detailed in the track overview paper[2]. The total number of submitted passages was $54,932$ with an average length of 32 tokens. The total number of tokens was $1,764,373$ with an average of $7,352$ per tweet. We also generated two reference runs based one the organizer's system made available to participants using 2013 and 2012 corpus respectively.

Informativeness results are presented in Table 6, with passage t-rels on the left and NPs t-rels on the right. Readability results are presented in Table 7. Note that the scores are divergences, and hence lower scores are better.

Both informativeness rankings in Table 6 are highly correlated, however discrepancies between the two rankings show that differences between top ranked runs rely on tokens outside NPs, mainly verbs since functional words are removed in the evaluation.

Table 7 reveals that readability of reference runs is low, meanwhile they are made of longer passages than average to ensure local syntax correctness.

Table 6. Informativeness results (official results are "with 2-gap").

Passage t-rels					NP t-rels				
Rank	Run	unigram	bigram	with 2-gap	Rank	Run	unigram	bigram	with 2-gap
1	ref2013	0.7050	0.7940	0.7960	1	ref2013	0.7468	0.8936	0.9237
2	ref2012	0.7528	0.8499	0.8516	2	ref2012	0.7784	0.9170	0.9393
3	361	0.7632	0.8689	0.8702	3	361	0.7903	0.9273	0.9461
4	360	0.7820	0.8925	0.8934	4	368	0.8088	0.9322	0.9486
5	368	0.8112	0.9066	0.9082	5	369	0.8090	0.9326	0.9489
6	369	0.8140	0.9098	0.9114	6	370	0.8131	0.9360	0.9513
7	359	0.8022	0.9120	0.9127	7	360	0.8104	0.9406	0.9553
8	370	0.8152	0.9137	0.9154	8	359	0.8227	0.9487	0.9613
9	356	0.8415	0.9696	0.9702	9	356	0.8477	0.9710	0.9751
10	357	0.8539	0.9700	0.9712	10	357	0.8593	0.9709	0.9752
11	364	0.8461	0.9697	0.9721	11	364	0.8628	0.9744	0.9807
12	358	0.8731	0.9832	0.9841	12	358	0.8816	0.9840	0.9864
13	362	0.8686	0.9828	0.9847	13	363	0.8840	0.9827	0.9870
14	363	0.8682	0.9825	0.9847	14	362	0.8849	0.9833	0.9876

Table 7. Readability results

Rank	Run	Relaxed (T)	Syntax (S)	Structure (A)	Average
1	358	0.948220	0.722796	0.721683	0.931005
2	356	0.952381	0.650917	0.703141	0.923958
3	357	0.948846	0.578212	0.713445	0.915750
4	362	0.836699	0.366561	0.608136	0.875917
5	363	0.836776	0.363954	0.611289	0.875500
6	364	0.880508	0.337197	0.639092	0.869167
7	359	0.930300	0.258563	0.535264	0.863375
8	360	0.925959	0.258658	0.588365	0.863274
9	361	0.932281	0.247883	0.501199	0.859749
10	ref2013	0.917378	0.259702	0.605203	0.857958
11	ref2012	0.913858	0.259584	0.606742	0.855583
12	369	0.912318	0.259539	0.549334	0.815625
13	368	0.908815	0.248981	0.565912	0.808750
14	370	0.901044	0.246893	0.538338	0.806958

Since reference runs are using the same system and index as the manual run used to build the t-rels, they tend to minimize the informativeness divergence with the reference. However, average divergence remains high pointing out that selecting the right passages in the restricted context of an entity, was more difficult than previous more generic tasks. Considering readability, the fact that reference runs are low ranked confirms that finding the right compromise between readability and informativeness remains the main difficulty of this task.

This year, the best participating system for informativeness used association rules. Since contextualization was restricted to some facet described by an entity, it could be that association rules helped to focus on this aspect.

The best participating system for readability used an advanced summarization systems that introduced minor changes in passages to improve readability. Changing the content of the passages was not allowed, however this tend to show that to deal with readability some rewriting is required. Moreover, since this year evaluation did not include a pool of passages from participants, systems that provided modified passages have been disadvantaged in informativeness evaluation.

4.5 Outlook

The discussion on next year's track is only starting, and there are links to related activities in other CLEF labs that need to be further explored.

5 Envoi

This complete our walk-through of INEX 2014. INEX 2014 focused on three tracks. The *Interactive Social Book Search Track* investigated user information seeking behavior when interacting with various sources of information, for realistic task scenarios, and how the user interface impacts search and the search experience. The *Social Book Search Track* investigated the relative value of authoritative metadata and user-generated content for search and recommendation using a test collection with data from Amazon and LibraryThing, and user profiles and personal catalogues. The *Tweet Contextualization Track* investigated tweet contextualization, helping a user to understand a tweet by providing him with a short background summary generated from relevant Wikipedia passages aggregated into a coherent summary (in collaboration with the RepLab Lab).

The INEX tracks cover various aspects of focused retrieval in a wide range of information retrieval tasks. This overview has only touched upon the various approaches applied to these tasks, and their effectiveness. The online proceedings of CLEF 2014 contains both the track overview papers [2, 4, 5], as well as the papers of the participating groups. The main result of INEX 2014, however, is a great number of test collections that can be used for future experiments, and the discussion amongst the participants that happens at the CLEF 2014 conference in Sheffield and throughout the year on the discussion lists.

References

[1] T. Beckers, N. Fuhr, N. Pharo, R. Nordlie, and K. N. Fachry. Overview and results of the INEX 2009 interactive track. In M. Lalmas, J. M. Jose, A. Rauber, F. Sebastiani, and I. Frommholz, editors, *ECDL*, volume 6273 of *Lecture Notes in Computer Science*, pages 409–412. Springer, 2010. ISBN 978-3-642-15463-8.

[2] P. Bellot, V. Moriceau, J. Mothe, E. Sanjuan, and X. Tannier. Overview of the INEX 2014 tweet contextualization track. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Labs and Workshops, Notebook Papers*, number 1180 in CEUR Workshop Proceedings, 2014. `http://ceur-ws.org/Vol-1181/`.

[3] P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of documentation*, 53(3): 225–250, 1997.

[4] M. Hall, H. Huurdeman, M. Koolen, M. Skov, and D. Walsh. Overview of the INEX 2014 interactive social book search track. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Labs and Workshops, Notebook Papers*, number 1180 in CEUR Workshop Proceedings, 2014. `http://ceur-ws.org/Vol-1181/`.

[5] M. Koolen, T. Bogers, G. Kazai, J. Kamps, and M. Preminger. Overview of the INEX 2014 social book search track. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Labs and Workshops, Notebook Papers*, number 1180 in CEUR Workshop Proceedings, 2014. `http://ceur-ws.org/Vol-1181/`.

[6] H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2009.

[7] V. Petras, T. Bogers, E. Toms, M. Hall, J. Savoy, P. Malak, A. Pawłowski, N. Ferro, and I. Masiero. Cultural heritage in CLEF (CHiC) 2013. In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *LNCS*, pages 192–211. Springer, 2013.

[8] M. Skov and P. Ingwersen. Exploring information seeking behaviour in a digital museum context. In *Proceedings of the second international symposium on Information interaction in context*, pages 110–115. ACM, 2008.

[9] E. G. Toms and M. M. Hall. The CHiC interactive task (CHiCi) at CLEF2013. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, 2013.