Overview of the INEX 2011 Books and Social Search Track

Marijn Koolen[1], Gabriella Kazai[2], Jaap Kamps[1], Antoine Doucet[3], and Monica Landoni[4]

[1] University of Amsterdam, Netherlands
{marijn.koolen,kamps}@uva.nl
[2] Microsoft Research, United Kingdom
v-gabkaz@microsoft.com
[3] University of Caen, France
doucet@info.unicaen.fr
[4] University of Lugano
monica.landoni@unisi.ch

Abstract. The goal of the INEX 2011 Books and Social Search Track is to evaluate approaches for supporting users in reading, searching, and navigating book metadata and full texts of digitized books. The investigation is focused around four tasks: 1) the Social Search for Best Books task aims at comparing traditional and user-generated book metadata for retrieval, 2) the Prove It task evaluates focused retrieval approaches for searching books, 3) the Structure Extraction task tests automatic techniques for deriving structure from OCR and layout information, and 4) the Active Reading task aims to explore suitable user interfaces for eBooks enabling reading, annotation, review, and summary across multiple books. We report on the setup and the results of the track.

1 Introduction

Prompted by the availability of large collections of digitized books, e.g., the Million Book project[5] and the Google Books Library project,[6] the Books and Social Search Track[7] was launched in 2007 with the aim to promote research into techniques for supporting users in searching, navigating and reading book metadata and full texts of digitized books. Toward this goal, the track provides opportunities to explore research questions around four areas:

- The relative value of professional and user-generated metadata for searching large collections of books,
- Information retrieval techniques for searching collections of digitized books,
- Mechanisms to increase accessibility to the contents of digitized books, and
- Users' interactions with eBooks and collections of digitized books.

[5] http://www.ulib.org/
[6] http://books.google.com/
[7] Until this year the Track was known as the Book Track.

Based around these main themes, the following four tasks were defined:

1. 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.
2. 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;
3. The *Structure Extraction* (SE) task aims at evaluating automatic techniques for deriving structure from OCR and building hyperlinked table of contents;
4. The *Active Reading task* (ART) aims to explore suitable user interfaces to read, annotate, review, and summarize multiple books.

In this paper, we report on the setup and the results of each of these tasks at INEX 2011. First, in Section 2, we give a brief summary of the participating organisations. The four task are described in detail in the following sections: the SB task in Section 3, the PI task in Section 4, the SE task in Section 5 and the ART in Section 6. We close in Section 7 with a summary and plans for INEX 2012.

2 Participating Organisations

A total of 47 organisations registered for the track (compared with 82 in 2010, 84 in 2009, 54 in 2008, and 27 in 2007). At the time of writing, we counted 10 active groups (compared with 16 in 2009, 15 in 2008, and 9 in 2007), see Table 1.[8]

3 The Social Search for Best Books Task

The goal of the Social Search for Best Books (SB) task is to evaluate the relative value of controlled book metadata, such as classification labels, subject headings and controlled keywords, versus user-generated or social metadata, such as tags, ratings and reviews, for retrieving the most relevant books for a given user request. Controlled metadata, such as the Library of Congress Classification and Subject Headings, is rigorously curated by experts in librarianship. It is used to index books to allow highly accurate retrieval from a large catalogue. However, it requires training and expertise to use effectively, both for indexing and for searching. On the other hand, social metadata, such as tags, are less rigorously defined and applied, and lack vocabulary control by design. However, such metadata is contributed directly by the users and may better reflect the terminology of everyday searchers. Clearly, both types of metadata have advantages and disadvantages. The task aims to investigate whether one is more suitable than the

[8] The last two groups participated in the SE task via ICDAR but did not register for INEX, hence have no ID.

Table 1. Active participants of the INEX 2011 Books and Social Search Track, the task they were active in, and number of contributed runs (SB = Social Search for Best Books, PI = Prove It, SE = Structure Extraction, ART = Active Reading Task)

ID	Institute	Tasks	Runs
4	University of Amsterdam	SB	6
7	Oslo University College	PI	15
18	Universitat Pompeu Fabra	SB	6
34	Nankai University	SE	4
50	University of Massachusettes	PI	6
54	Royal School of Library and Information Science	SB	4
62	University of Avignon	SB	6
113	University of Caen	SE	3
	Microsoft Development Center Serbia	SE	1
	Xerox Research Centre Europe	SE	2

other to support different types of search requests or how they may be fruitfully combined.

The SB task aims to address the following research questions:

- How can a system take full advantage of the available metadata for searching in an online book catalogue?
- What is the relative value of social and controlled book metadata for book search?
- How does the different nature of these metadata descriptions affect retrieval performance for different topic types and genres?

3.1 Scenario

The scenario is that of a user turning to Amazon Books and LibraryThing to search for books they want to read, buy or add to their personal catalogue. Both services host large collaborative book catalogues that may be used to locate books of interest.

On LibraryThing, users can catalogue the books they read, manually index them by assigning tags, and write reviews for others to read. Users can also post messages on a discussion forum asking for help in finding new, fun, interesting, or relevant books to read. The forums allow users to tap into the collective bibliographic knowledge of hundreds of thousands of book enthusiasts. On Amazon, users can read and write book reviews and browse to similar books based on links such as "customers who bought this book also bought... ".

Users can search online book collections with different intentions. They can search for specific books of which they know all the relevant details with the intention to obtain them (buy, download, print). In other cases, they search for a specific book of which they do not know those details, with the intention of

identifying that book and find certain information about it. Another possibility is that they are not looking for a specific book, but hope to discover one or more books meeting some criteria. These criteria can be related to subject, author, genre, edition, work, series or some other aspect, but also more serendipitously, such as books that merely look interesting or fun to read.

Although book metadata can often be used for browsing, this task assumes a user issues a query to a retrieval system, which returns a (ranked) list of book records as results. This query can be a number of keywords, but also one or more book records as positive or negative examples. We assume the user inspects the results list starting from the top and works her way down until she has either satisfied her information need or gives up. The retrieval system is expected to order results by relevance to the user's information need.

3.2 Task description

The SB task is to reply to a user's request that has been posted on the LibraryThing forums (see Section 3.5) by returning a list of recommended books. The books must be selected from a corpus that consists a collection of book metadata extracted from Amazon Books and LibraryThing, extended with associated records from library catalogues of the Library of Congress and the British Library (see the next section). The collection includes both curated and social metadata. User requests vary from asking for books on a particular genre, looking for books on a particular topic or period or books by a given author. The level of detail also varies, from a brief statement to detailed descriptions of what the user is looking for. Some requests include examples of the kinds of books that are sought by the user, asking for similar books. Other requests list examples of known books that are related to the topic but are specifically of no interest. The challenge is to develop a retrieval method that can cope with such diverse requests. Participants of the SB task are provided with a set of book search requests and are asked to submit the results returned by their systems as ranked lists.

3.3 Submissions

We want to evaluate the book ranking of retrieval systems, specifically the top ranks. We adopt the submission format of TREC, with a separate line for each retrieval result, consisting of six columns:

1. topic_id: the topic number, which is based on the LibraryThing forum thread number.
2. Q0: the query number. Unused, so should always be Q0.
3. isbn: the ISBN of the book, which corresponds to the file name of the book description.
4. rank: the rank at which the document is retrieved.
5. rsv: retrieval status value, in the form of a score. For evaluation, results are ordered by descending score.

6. run_id: a code to identifying the participating group and the run.

Participants are allowed to submit up to six runs, of which at least one should use only the *title* field of the topic statements (the topic format is described in Section 3.5). For the other five runs, participants could use any field in the topic statement.

3.4 Data

To study the relative value of social and controlled metadata for book search, we need a large collection of book records that contains controlled subject headings and classification codes as well as social descriptions such as tags and reviews, for a set of books that is representative of what readers are searching for. We use the Amazon/LibraryThing corpus crawled by the University of Duisburg-Essen for the INEX Interactive Track [1].

The collection consists of 2.8 million book records from Amazon, extended with social metadata from LibraryThing. This set represents the books available through Amazon. These records contain title information as well as a Dewey Decimal Classification (DDC) code and category and subject information supplied by Amazon. From a sample of Amazon records we noticed the subject descriptors to be noisy, with many inappropriately assigned descriptors that seem unrelated to the books to which they have been assigned.

Each book is identified by ISBN. Since different editions of the same work have different ISBNs, there can be multiple records for a single intellectual work. The corpus consists of a collection of 2.8 million records from Amazon Books and LibraryThing.com. See https://inex.mmci.uni-saarland.de/data/nd-agreements.jsp for information on how to get access to this collection. Each book record is an XML file with fields like ¡isbn¿, ¡title¿, ¡author¿, ¡publisher¿, ¡dimensions¿, ¡numberofpage¿ and ¡publicationdate¿. Curated metadata comes in the form of a Dewey Decimal Classification in the ¡dewey¿ field, Amazon subject headings are stored in the ¡subject¿ field, and Amazon category labels can be found in the ¡browseNode¿ fields. The social metadata from Amazon and LibraryThing is stored in the ¡tag¿, ¡rating¿, and ¡review¿ fields. The full list of fields is shown in Table 2.

How many of the book records have curated metadata? There is a DDC code for 61% of the descriptions and 57% of the collection has at least one subject heading. The classification codes and subject headings cover the majority of records in the collection.

More than 1.2 million descriptions (43%) have at least one review and 82% of the collection has at least one LibraryThing tag.

The distribution of books over the Amazon subject categories shows that *Literature*, *History*, *Professional and Technical* and *Religion* are some of the largest categories (see Table 3). There are also administrative categories related to sales, edition (paperback, hardcover) and others, but we show only the genre-related categories. If we look at the distribution over DDC codes (showing only the main classes in Table 4), we see a somewhat different distribution. *Literature*

Table 2. A list of all element names in the book descriptions

tag name			
book	similarproducts	title	imagecategory
dimensions	tags	edition	name
reviews	isbn	dewey	role
editorialreviews	ean	creator	blurber
images	binding	review	dedication
creators	label	rating	epigraph
blurbers	listprice	authorid	firstwordsitem
dedications	manufacturer	totalvotes	lastwordsitem
epigraphs	numberofpages	helpfulvotes	quotation
firstwords	publisher	date	seriesitem
lastwords	height	summary	award
quotations	width	editorialreview	browseNode
series	length	content	character
awards	weight	source	place
browseNodes	readinglevel	image	subject
characters	releasedate	imageCategories	similarproduct
places	publicationdate	url	tag
subjects	studio	data	

Table 3. Amazon category distribution (in percentages)

Category	%	**Category**	%
Non-fiction	20	Science	7
Literature and fiction	20	Fiction	7
Children	14	Literature	7
History	13	Christianity	7
Reference	11	Health, Mind and Body	6
Professional and Technical.	11	Arts and Photography	5
Religion and Spirituality	10	Business and Investing	5
Social science	10	Biography and Memoirs	5

Table 4. Distribution over DDC codes (in percentages)

DDC main class	%
Computer science, information and general works	4
Philosophy and psychology	4
Religion	8
Social sciences	16
Language	2
Science (including mathematics)	5
Technology and applied Science	13
Arts and recreation	13
Literature	25
History, geography, and biography	11

is still the largest class, but is followed by *Social sciences*, *Arts and recreation*, *Technology*, then *History* and *Religion*. Note that a book has only one DDC code—it can only have one physical location on a library shelf—but can have multiple Amazon categories, which could explain the difference in distribution. Note also that all but 296 books in the collection have at least one Amazon category, while only 61% of the records have DDC codes.

3.5 Information needs

LibraryThing users discuss their books in the discussion forums. Many of the topic threads are started with a request from a member for interesting, fun new books to read. They describe what they are looking for, give examples of what they like and do not like, indicate which books they already know and ask other members for recommendations. Other members often reply with links to works catalogued on LibraryThing, which have direct links to the corresponding records on Amazon. These requests for recommendation are natural expressions of information needs for a large collection of online book records. We aim to evaluate the SB task using a selection of these forum topics.

The books suggested by members in replies to the initial message are collected in a list on the side of the topic thread (see Figure 1). A technique called *touchstone* can be used by members to easily identify books they mention in the topic thread, giving other readers of the thread direct access to a book record on LibraryThing, with associated ISBNs and links to Amazon. We use these suggested books as initial relevance judgements for evaluation. Some of these touchstones identify an incorrect book, and suggested books may not always be what the topic creator asked for, but merely be mentioned as a negative example or for some other reason. From this it is clear that the collected list of suggested books can contain false positives and is probably incomplete as not all relevant books will be suggested (false negatives), so may not be appropriate for reliable evaluation. We discuss this in more detail in Section 3.7. We first describe how

Fig. 1. A topic thread in LibraryThing, with suggested books listed on the right hand side.

we created a large set of topics, then analyse what type of topics we ended up with and how suitable they for this task.

Topic analysis We crawled 18,427 topic threads from 1,560 discussion groups. From these, we extracted 943 topics where the initial message contains a request for book suggestions. Each topic has a title and is associated with a group on the discussion forums. For instance, topic 99309 in Figure 1 has title *Politics of Multiculturalism Recommendations?* and was posted in the group *Political Philosophy*. Not all titles are good descriptions of the information need expressed in the initial message. To identify which of these 943 topics have good descriptive titles, we used the titles as queries and retrieved records from the Amazon/LibraryThing collection and evaluated them using the suggested books collected through the touchstones. We selected all topics for which at least 50% of the suggested books were returned in the top 1000 results and manually labelled them with information about topic type, genre and specificity and extracted positive and negatives example books and authors mentioned in the initial message. Some topics had very vague requests or relied on external source to derive the information need (such as recommendations of books listed on certain web page), leaving 211 topics in the official test topic set from 122 different discussion groups.

To illustrate how we marked up the topics, we show topic 99309 from Figure 1 as an example:

```
<topic id="99309">
  <title>Politics of Multiculturalism</title>
  <group>Political Philosophy</group>
```

```
  <narrative>I'm new, and would appreciate any recommended reading on the
    politics of multiculturalism. <author>Parekh</author>'s
    <work id="164382"> Rethinking Multiculturalism: Cultural Diversity and
    Political Theory</work> (which I just finished) in the end left me un-
    convinced, 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 skepti-
    cism, but still... Anyway, I've read <author>Sen</author>, <author>
    Rawls</author>, <author>Habermas</author>, and <author>Nussbaum
    </author>, 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>
  <type>subject</type>
  <genre>politics</genre>
  <specificity>narrow</specificity>
  <similar>
    <work id="164382">
      <isbn>0333608828</isbn>
      <isbn>0674004361</isbn>
      <isbn>1403944539</isbn>
      <isbn>0674009959</isbn>
    </work>
    <author>Parekh</author>
    <author>Sen</author>
    <author>Rawls</author>
    <author>Habermas</author>
    <author>Nussbaum</author>
  </similar>
  <dissimilar><dissimilar>
</topic>
```

The distribution over topic type is shown on the left side of Table 5. The majority of topics have subject-related book requests. For instance, the topic in Figure 1 is a subject-related request, asking for books about politics and multiculturalism. Most requests (64%) are subject-related, followed by author-related (15%), then series (5%), genre (4%), edition and known-item (both 3%). Some topics can be classified with 2 types, such as subject and genre. For instance, in one topic thread, the topic creator asks for biographies of people with eating disorders. In this case, the subject is *people with eating disorders* and the genre is *biography*. The topic set covers a broad range of topic types, but for work- and language-related topics the numbers are too small to be representative. We will conduct a more extensive study of the topics to see if this distribution is representative or whether our selection method has introduced some bias.

Next, we classified topics by genre, roughly based on the main classes of the LCC and DDC (see right side of Table 5), using separate classes for philosophy and religion (similar to DDC, while LCC combines them in one main class). The two most requested genres are literature (42%, mainly prose and some poetry), and history (28%). We only show the 12 most frequent classes. There are more

Table 5. Distribution of topic types and genres

Type	Freq.	Genre	Freq.
subject	134	literature	89
author	32	history	60
series	10	biography	24
genre	8	military	16
edition	7	religion	16
known-item	7	technology	14
subject & genre	7	science	11
work	2	education	8
genre & work	1	politics	4
subject & author	1	philosophy	4
language	1	medicine	3
author & genre	1	geography	3

main classes represented by the topics, such as law, psychology and genealogy, but they only represent one or two topics each. If we compare this distribution with the Amazon category and DDC distributions in Tables 3 and 4, we see that military books are more popular among LibraryThing forum users than is represented by the Amazon book corpus, while social science is less popular. Literature, history, religion, technology are large class in both the book corpus and the topic set. The topic set is a reasonable reflection of the genre distribution of the books in the Amazon/LibraryThing collection.

Furthermore, we added labels for specificity. The specificity of a topic is somewhat subjective and we based it on a rough estimation of the number of relevant books. It is difficult to come up with a clear threshold between broad and narrow, and equally hard to estimate how many books would be relevant. Broad topics have requests such as recommendations within a particular genre ("please recommend good science fiction books."), for which thousands of books could be considered relevant. The topic in Figure 1 is an example of a narrow topic. There are 177 topics labelled as narrow (84%) and 34 topics as broad (16%). We also labelled books mentioned in the initial message as either positive or negative examples of what the user is looking for. There are 58 topics with positive examples (27%) and 9 topics with negative examples (4%). These topics could be used as query-by-example topics, or maybe even for recommendation. The examples add further detail to the expressed information need and increase the realism of the topic set.

We think this topic set is representative of book information needs and expect it to be suitable for evaluating book retrieval techniques. We note that the titles and messages of the topic threads may be different from what these users would submit as queries to a book search system such as Amazon, LibraryThing, the Library of Congress or the British Library. Our topic selection method is an attempt to identify topics where the topic title describes the information need. In the first year of the task, we ask the participants to generate queries from

Table 6. Statistics on the number of recommended books for the 211 topics from the LT discussion groups

# rel./topic	# topics	min.	max.	median	mean	std. dev.
All	211	1	79	7	11.3	12.5
Fiction	89	1	79	10	16.0	15.8
Non-fiction	132	1	44	6	8.3	8.3
Subject	142	1	68	6	9.6	10.0
Author	34	1	79	10	15.9	17.6
Genre	16	1	68	7	13.3	16.4

the title and initial message of each topic. In the future, we could approach the topic creators on LibraryThing and ask them to supply queries or set up a crowdsourcing task where participants provide queries while searching the Amazon/LibraryThing collection for relevant books.

Touchstone Recommendations as Judgements We use the recommended books for a topic as relevance judgements for evaluation. Each book in the Touchstone list is considered relevant. How many books are recommended to LT members requesting recommendations in the discussion groups? Are other members compiling exhaustive lists of possibly interesting books or do they only suggest a small number of the best available books? Statistics on the number of books recommended for the 211 topics are given in Table 6.

The number of relevant books per topic ranges between 1 and 79 with a mean of 11.3. The median is somewhat lower (7), indicating that most of the topics have a small number of recommended books. The topics requesting fiction books have more relevant books (16 on average) than the topics requesting non-fiction (8.3 on average). Perhaps this is because there is both more fiction in the collection and more fiction related topics in the topic set. The latter point suggests that fiction is more popular among LT members, such that requests for books get more responses. The breakdown over topic types *Subject*, *Author* and *Genre* shows that subject related topics have fewer suggested books than author and genre related topics. This is probably related to the distinction between fiction and non-fiction. Most of the *Subject* topics are also *Non-fiction* topics, which have fewer recommended books than *Fiction* books.

ISBNs and intellectual works
Each record in the collection corresponds to an ISBN, and each ISBN corresponds to a particular intellectual work. An intellectual work can have different editions, each with their own ISBN. The ISBN-to-work relation is a many-to-one relation. In many cases, we assume the user is not interested in all the different editions, but in different intellectual works. For evaluation we collapse multiple ISBN to a single work. The highest ranked ISBN is evaluated and all lower ranked ISBNs ignored. Although some of the topics on LibraryThing are requests to recommend a particular edition of a work—in which case the distinction between different

ISBNs for the same work are important—we leave them out of the relevance assessment phase for this year to make evaluation easier.

However, one problem remains. Mapping ISBNs of different editions to a single work is not trivial. Different editions may have different titles and even have different authors (some editions have a foreword by another author, or a translator, while others have not), so detecting which ISBNs actually represent the same work is a challenge. We solve this problem by using mappings made by the collective work of LibraryThing members. LT members can indicate that two books with different ISBNs are actually different manifestations of the same intellectual work. Each intellectual work on LibraryThing has a unique work ID, and the mappings from ISBNs to work IDs is made available by LibraryThing.[9]

However, the mappings are not complete and might contain errors. Furthermore, the mappings form a many-to-many relationship, as two people with the same edition of a book might independently create a new book page, each with a unique work ID. It takes time for members to discover such cases and merge the two work IDs, which means that at time, some ISBNs map to multiple work IDs. LibraryThing can detect such cases but, to avoid making mistakes, leaves it to members to merge them. The fraction of works with multiple ISBNs is small so we expect this problem to have a negligible impact on evaluation.

3.6 Crowdsourcing Judgements on Relevance and Recommendation

Members recommend books they have read or that they know about. This may be only a fraction of all the books that meet the criteria of the request. The list of recommended books in a topic thread may therefore be an incomplete list of appropriate books. Retrieval systems can retrieve many relevant books that are not recommended in the thread. On the other hand, LT members might leave out certain relevant books on purpose because they consider these books inferior to the books they do suggest.

To investigate this issue we ran an experiment on Amazon Mechanical Turk, where we asked workers to judge the relevance and make recommendations for books based on the descriptions from the Amazon/LT collection. For the PI task last year we found that relevance judgements for digitised book pages from AMT give reliable system rankings [5]. We expect that judging the relevance of an Amazon record given a narrative from the LibraryThing discussion forum has a lower cognitive load for workers, and with appropriate quality-control measures built-in, we expect AMT judgements on book metadata to be useful for reliable evaluation as well. An alternative or complement is to ask task participants to make judgements.

We pooled the top 10 results of all official runs for 24 topics and had each book judged by 3 workers. We explicitly asked workers to first judge the book on topical relevance and with a separate question asked them to indicate whether they would also recommend it as one the best books on the requested topic.

Topic selection

[9] See: http://www.librarything.com/feeds/thingISBN.xml.gz

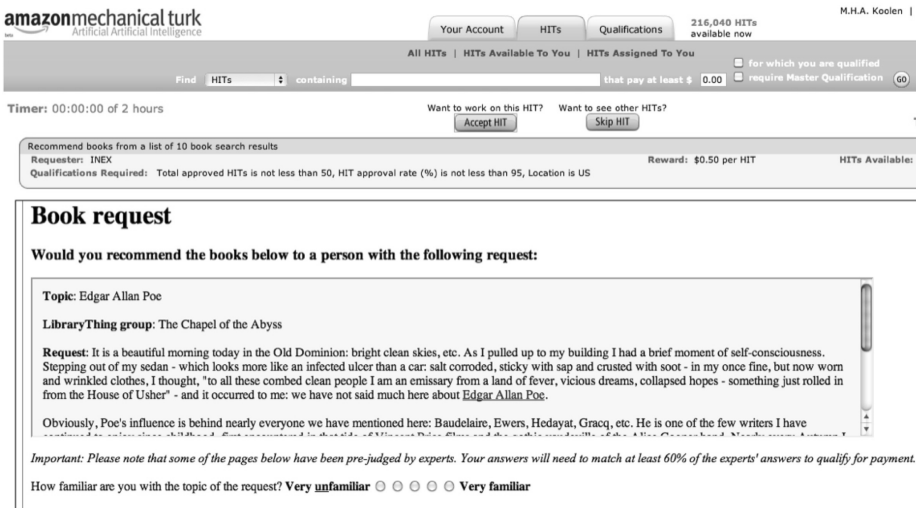

Fig. 2. Snapshot of the AMT book request.

For the Mechanical Turk judgements, we 24 topics from the set of 211, 12 fiction and 12 non-fiction. We selected the following 12 fiction topics: 17299, 25621, 26143, 28197, 30061, 31874, 40769, 74433, 84865, 94888, 92178 and 106721. We selected the following 12 non-fiction topics: 3963, 12134, 14359, 51583, 65140, 83439, 95533, 98106, 100674, 101766, 107464 and 110593.

Pooling
We pooled the top 10 results per topic of all 22 submitted runs. If the resulting pool was smaller than 100 books, we continued the round-robin pooling until each pool contained at least 100 books.

Generating HITs
Each HIT contains 10 books, with at least one book that was recommended in the topic thread on the LibraryThing discussion group for validation. In total, 269 HITs were generated, and each HIT was assigned to 3 workers, who got paid $0.50 per HIT. With a 10% fee charged by Amazon per HIT, the total cost was $269 * 3 * \$0.50 * 1.1 = \$443.85$.

HIT design
The design of the HIT is illustrated in Figures 2, 3, 4 and 5. The HIT starts with short instructions explaining what the task is and what the goal of the task is, after which the request is shown (see Figure 2). After the request, worker get a list of 10 book questionnaires, with each questionnaire containing frame with official metadata (Figure 3), user-generated metadata (Figure 4) and a list of questions (Figure 5). The official metadata consists of the title information, publisher information and the Amazon categories, subject headings and classification information. The user-generated metadata consists of user reviews and ratings from Amazon and user tags from LibraryThing.

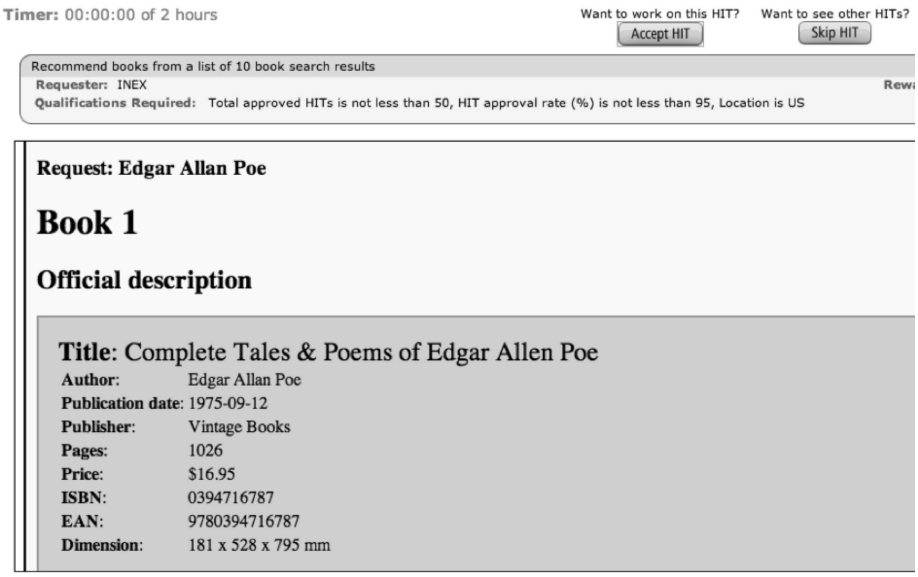

Fig. 3. Snapshot of the AMT design for the official description.

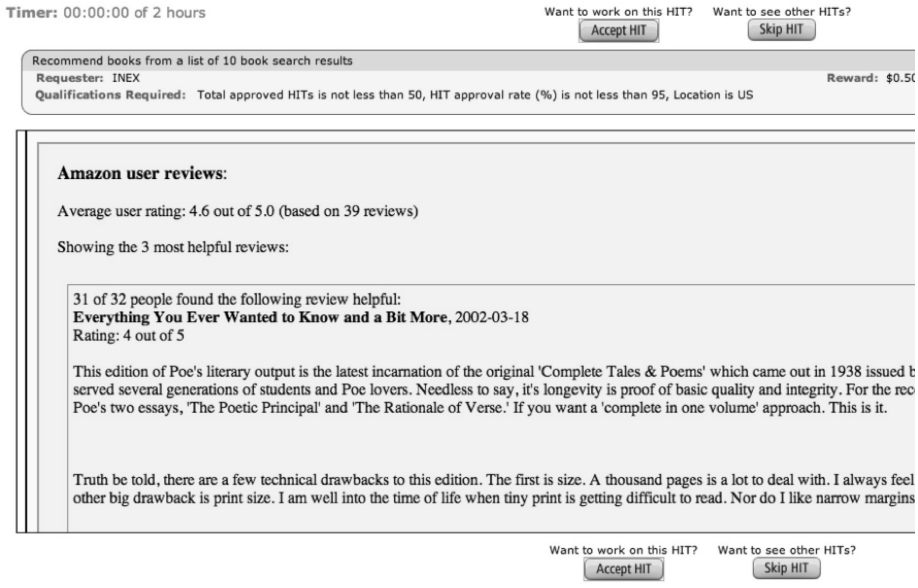

Fig. 4. Snapshot of the AMT design for the user-generated description.

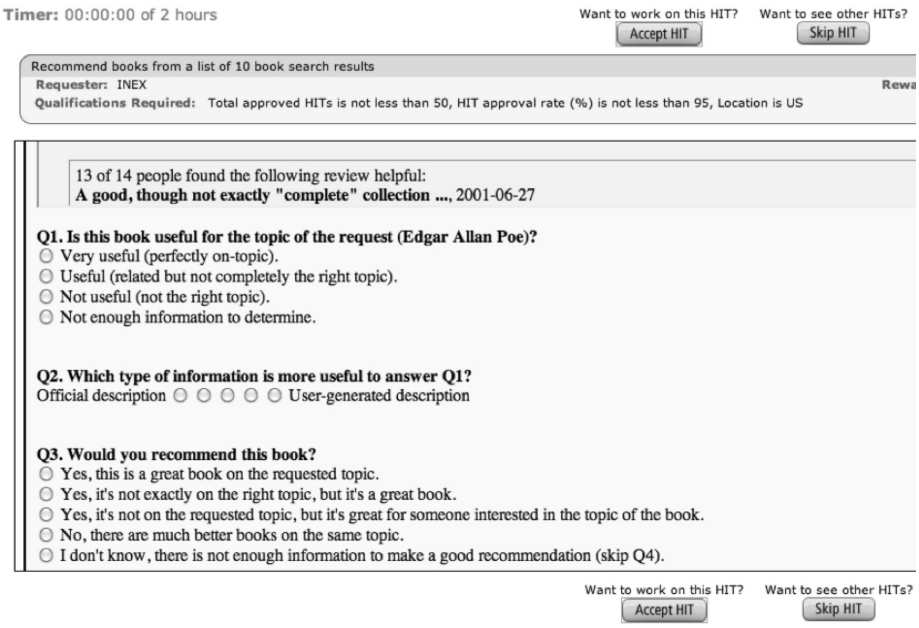

Fig. 5. Snapshot of the AMT questionnaire design.

The questionnaire has 5 questions:

- **Q1. Is this book useful for the topic of the request?** Here workers can choose between
  - *perfectly on-topic*,
  - *related but not completely the right topic*,
  - *not the right topic* and
  - *not enough information*.
- **Q2. Which type of information is more useful to answer Q1?** Here workers have to indicate whether the official or user-generated metadata is more useful to determine relevance.
- **Q3. Would you recommend this book?** Here workers can choose between
  - *great book on the requested topic*,
  - *not exactly on the right topic, but it's a great book*,
  - *not on the requested topic, but it's great for someone interested in the topic of the book*,
  - *there are much better books on the same topic,* and
  - *not enough information to make a good recommendation*.
- **Q4. Which type of information is more useful to answer Q3?** Here workers have to indicate whether the official or user-generated metadata is more useful to base their recommendation on.

Table 7. Statistics on the number of recommended books for the 211 topics from the LT discussion groups

# rel./topic	# topics	min.	max.	median	mean	std. dev.
LT all	211	1	79	7	11.3	12.5
LT Fiction	89	1	79	10	16.0	15.8
LT Non-fiction	132	1	44	6	8.3	8.3
LT (24 AMT topics)	24	2	79	7	15.7	19.3
AMT all	24	4	56	25	25.0	12.7
AMT fiction	12	4	30	25	22.8	10.8
AMT non-fiction	12	4	56	29	27.3	13.7

– **Q5. Please type the most useful tag (in your opinion) from the LibraryThing tags in the User-generated description.** Here workers had to pick one of the LibraryThing user tags as the most useful, or tick the box *or tick here if there are no tags for this book* when the user-generated metadata has no tags.

There was also an optional comments field per book.

Agreement
What is the agreement among workers? We compute the pairwise agreement on relevance among workers per HIT in three different ways. The most strict agreement distinguishes between the four possible answers: 1) *Perfectly on-topic*, 2) *related but not perfect*, 3) *not the right topic* and 4) *not enough information*. In this case agreement is 0.54. If we consider only answer 1 as relevant and merge answers 2 and 3 (related means not relevant), agreement is 0.63. If we also take answer 4 to mean non-relevant (merging 2, 3 and 4, giving binary judgements), agreement is 0.68.

Recall that each HIT has at least one book that is recommended on the LT discussion thread. The average agreement between workers and forum members is 0.52. That is, on average, each worker considered 52% of the books recommended on LT as *perfectly on-topic*. We turn the AMT relevance data from multiple workers into binary relevance judgements per book using majority vote. We only consider the *perfectly on-topic* category as relevant and map the other categories to non-relevant. For most books we have 3 votes, which always leads to a majority. Some books occur in multiple HITs because they are added as known relevant books from the LT forums. If there are fewer recommended books in the LT forum than there are HITs, some books have to be included in multiple HITs. Books with judgements from an even number of workers could have tied votes. In these cases we use the fact that the book was recommended on the LT topic thread as the deciding vote and label the book as relevant.

How does the relevance distribution of the AMT judgements compare to the relevance judgements from the LT discussion groups? We compare the AMT relevance judgements with the recommendations from LT in Table 7. The fiction topics have more LT recommendations than the non-fiction, but fewer relevant

Table 8. Evaluation results for the official submissions using the LT relevance judgements of all 211 topics. Best scores are in bold.

Run	nDCG@10	P@10	MRR	MAP
p4-inex2011SB.xml_social.fb.10.50	**0.3101**	**0.2071**	**0.4811**	**0.2283**
p54-run4.all-topic-fields.reviews-split.combSUM	0.2991	0.1991	0.4731	0.1945
p4-inex2011SB.xml_social	0.2913	0.1910	0.4661	0.2115
p54-run2.all-topic-fields.all-doc-fields	0.2843	0.1910	0.4567	0.2035
p62.recommandation	0.2710	0.1900	0.4250	0.1770
p62.sdm-reviews-combine	0.2618	0.1749	0.4361	0.1755
p18.UPF_QE_group_BTT02	0.1531	0.0995	0.2478	0.1223
p18.UPF_QE_genregroup_BTT02	0.1327	0.0934	0.2283	0.1001

books according to the AMT workers. This might be a sign that, without having read the book, judging the relevance of fiction books is harder than that of non-fiction books. For fiction there is often more to the utility of a book (whether it is interesting and/or fun) than the subject and genre information provided by book metadata. Or perhaps the relevance of fiction books is not harder to judge, but fiction is less readily considered relevant. For non-fiction information needs, the subject of a book may be one of the main aspects on which the relevance of the book is based. For fiction information needs, the subject of a book might play no role in determine its relevance. Another explanation might that the judgements pools based on the official runs might be better for non-fiction topics than for fiction topics.

3.7 Evaluation

For some topics, relevance may be both trivial and complex. Consider a topic where a user asks for good historical fiction books. The suggestions from the LT members will depend on their ideas of what are good historical fiction books. From the metadata alone it is hard to make this judgement. Should all historical fiction books be considered relevant, or only the ones suggested by the LT members? Or should relevance be graded?

For now, we will use a one-dimensional relevance scale, but like to explore alternatives in the future. One way would be to distinguish between books that a user considers as interesting options to read next and the actual book or books she decides to obtain and read. This roughly corresponds to the distinction between the library objective of helping to *find or locate* relevant items and the objective of helping to *choose* which of the relevant items to access [7].

We first show the results for the 211 topics and associated relevance judgements from the LT forums in Table 8. The best SB run (nDCG@10=0.3101) was submitted by the University of Amsterdam (p4-inex2011SB.xml_social.fb.10.50), which uses pseudo relevance feedback on an index with only reviews and tags in addition with the basic title information.

Table 9. Evaluation results for the official submissions using the AMT relevance judgements. Best scores are in bold.

Run	nDCG@10	P@10	MRR	MAP
p62.baseline-sdm	**0.6092**	**0.5875**	0.7794	0.3896
p4-inex2011SB.xml_amazon	0.6055	0.5792	0.7940	0.3500
p62.baseline-tags-browsenode	0.6012	0.5708	0.7779	**0.3996**
p4-inex2011SB.xml_full	0.6011	0.5708	0.7798	0.3818
p54-run2.all-topic-fields.all-doc-fields	0.5415	0.4625	**0.8535**	0.3223
p54-run3.title.reviews-split.combSUM	0.5207	0.4708	0.7779	0.2515
p18.UPF_base_BTT02	0.4718	0.4750	0.6276	0.3269
p18.UPF_QE_group_BTT02	0.4546	0.4417	0.6128	0.3061

What is surprising is that some systems score high on MRR while the number of forum suggestions is small and not based on top-k pooling. With only a median of 7 suggested books per topic in a collection of 2.8 million books, the forum suggestions may well be highly incomplete. That is, there might be hundreds of other books that would have made equally good suggestions. If that were the case, it would be difficult for retrieval systems to obtain a high score, as they would likely place different books in the top ranks than the forum members. If there are a hundred relevant books in the collection and the forum members randomly picked 7 of them as suggestions, the probability that a retrieval system will rank several of those 7 in the top 10 is small. Of course, this could happen for a single topic, but an average MRR of 0.4811 over 211 topics would be extremely unlikely. This suggests that the forum suggestions are not drawn from a much larger set of equally relevant books, but form a more or less complete set of the best or most popular books for the requested topic.

Next we show the results for the 24 topics selected for the AMT experiment and associated relevance judgements in Table 9. The best SB run (nDCG@10=0.6092) was submitted by the University of Avignon (p62-baseline-sdm). The most striking difference with the LT forum judgements is that here the scores for all runs are much higher. There are at least three possible explanations for this. First, the AMT judgements are based on the top 10 results of all runs, meaning all top 10 results of each run is judged, whereas many top ranked documents are not covered by the LT forum judgements. Second, the AMT judgements are explicitly based on topical relevance, whereas the LT forum judgements are probably more like recommendations, where users only suggest the best books on a topic and often only books they know about or have read. The high scores of the submitted runs indicates that systems are good a finding topically relevant books. The third possible explanation is that the two evaluations are based on different topic sets. The LT forum evaluation is based on 211 topics, while the AMT evaluation is based on a subset of 24 topics.

To rule out that last explanation, we also evaluated the submitted runs using the LT forum judgements only on the subset of 24 topics selected for the AMT experiment. The results for this are shown in Table 10. The topic set has little

Table 10. Evaluation results for the official submissions using the LT relevance judgements for the 24 topics used in AMT. Best scores are in bold.

Run	ndcg@10	P@10	MRR	MAP
p4-inex2011SB.xml_social.fb.10.50	**0.3039**	**0.2120**	**0.5339**	0.1994
p54-run2.all-topic-fields.all-doc-fields	0.2977	0.1940	0.5225	**0.2113**
p4-inex2011SB.xml_social	0.2868	0.1980	0.5062	0.1873
p54-run4.all-topic-fields.reviews-split.combSUM	0.2601	0.1940	0.4758	0.1515
p62.recommandation	0.2309	0.1720	0.4126	0.1415
p62.sdm-reviews-combine	0.2080	0.1500	0.4048	0.1352
p18.UPF_QE_group_BTT02	0.1073	0.0720	0.2133	0.0850
p18.UPF_QE_genregroup_BTT02	0.0984	0.0660	0.1956	0.0743

Table 11. Evaluation results using the LT recommendation Qrels across fiction and non-fiction topics.

	nDCG@10		
Run	**All**	**Fiction**	**Non-fiction**
p4-inex2011SB.xml_social.fb.10.50	**0.3101**	**0.3469**	0.2896
p54-run4.all-topic-fields.reviews-split.combSUM	0.2991	0.3062	**0.2908**
p4-inex2011SB.xml_social	0.2913	0.3157	0.2783
p54-run2.all-topic-fields.all-doc-fields	0.2843	0.3145	0.2627
p62.recommandation	0.2710	0.2779	0.2694
p62.sdm-reviews-combine	0.2618	0.2680	0.2609
p18.UPF_QE_group_BTT02	0.1531	0.1505	0.1533
p18.UPF_QE_genregroup_BTT02	0.1327	0.1474	0.1238

impact, as the results for the subset of 24 topics are very similar to the results for the 211 topics. This is a first indication that the LT forum test collection is robust with respect to topic selection. It also suggests that the LT forum and AMT judgements reflect different tasks. The latter is the more traditional topical relevance task, while the former is closer to recommendation. We are still in the process of analysing the rest of the AMT data to establish to what extent the LT forum suggestions reflect relevance and recommendation tasks.

Recall that we added genre labels to all the topics. We divide the topics into two sets, one with fiction related topics and one with non-fiction related topics. All the topics with the label *literature* are considered fiction related. All other topics are considered non-fiction topics. Table 11 shows the nDCG@10 results over the topics sets split over topic genre. Most systems perform slightly better on the fiction topics than on the non-fiction topics. One reason might be that more books are suggested for fiction-related topics (see Table 7). Another reason might be that fiction books are more popular and therefore have more detailed descriptions in the form of tags and reviews and are easier to retrieve and rank.

We also split the 211 topics over topic types. The most frequent topic types are *subject* (books on a particular subject), *author* (books by a particular author)

Table 12. Evaluation results using the LT recommendation Qrels across fiction and non-fiction topics.

Run	nDCG@10			
	All	**Subject**	**Author**	**Genre**
p4-inex2011SB.xml_social.fb.10.50	**0.3101**	0.2644	**0.4645**	0.1466
p54-run4.all-topic-fields.reviews-split.combSUM	0.2991	**0.2658**	0.4368	0.1905
p4-inex2011SB.xml_social	0.2913	0.2575	0.4006	0.1556
p54-run2.all-topic-fields.all-doc-fields	0.2843	0.2435	0.4002	**0.2029**
p62.recommandation	0.2710	0.2411	0.3866	0.1248
p62.sdm-reviews-combine	0.2618	0.2386	0.3686	0.1250
p18.UPF_QE_group_BTT02	0.1531	0.1116	0.2331	0.0401
p18.UPF_QE_genregroup_BTT02	0.1327	0.1021	0.1913	0.0566

and *genre* (books in a particular genre). The nDCG@10 results over topic types are shown in Table 12. The general pattern is that author topics are easier than subject and genre topics, and subject topics are easier than genre topics. This is not surprising, given that author names are often highly specific which makes them good retrieval cues. With only a small set of books matching the author name, it is not hard to retrieve the books suggested by forum members. Subject descriptions are less specific and target a larger set of books, making it harder to single out the suggested books from other books on the same subject. Finally, genre labels are even less specific and vague at best. Forum members argue over different definitions science fiction or whether a book is fiction or non-fiction. The set of books belonging to a genre can also be very large—thousands or ten of thousands of books—such that forum members disagree over what the best suggestions are. These topics may be harder than other topic types, and as a result, IR systems perform poorly on these topics.

To sum up, the forum suggestions represent a different task from traditional topical relevance search. They introduce no pooling bias, but the fact that some systems score high on early precision indicate the suggestions are relatively complete. The high scores for the AMT judgements indicates that many systems are capable of finding topically relevant books, which further indicates that the book suggestions represent an interesting and realistic new task. There are no big differences between book requests for fiction and non-fiction, and author-related topics are easier than subject-related topics, which in turn are easier than genre-related topics.

3.8 Discussion

Relevance or recommendation?
Readers may not only base their judgement on the topical relevance—is this book a historical fiction book—but also on their personal taste. Reading a book is often not just about relevant content, but about interesting, fun or engaging content. Relevance in book search might require different dimensions of graded

judgements. The topical dimension (how topically relevant is this book?) is separate from the interestingness dimension (how interesting/engaging is this book?) Many topic creators ask for recommendations, and want others to explain their suggestions, so that they can better gauge how a book fits their taste.

So far, we only use the suggestions in the forum discussions as binary relevance judgements. However, the forum discussions contain more information than that. Some books are suggested by multiple forum members, and some books receive a negative recommendation. On top of that, the suggestions come from other members than the topic creator, and might not coincide with her actual interest.

We will report on the analysis of the AMT questionnaire data separately, but preliminary results suggest that workers treat topical relevance and recommendation similarly. When they consider a book topically relevant, they almost always recommend it as well, and do not recommend it when it is not relevant, even though there is an answer category for books that are on a different topic but are very good books for that topic. This might be the case because we asked workers to judge topical relevance and recommendation in the same questionnaire. Because the questions about recommendation were framed in the context of a specific book request, workers may have interpreted recommendation in terms of that request. Also, most books have favourable reviews, which makes it harder to distinguish between books other than to look at their relation to the requested topic. For books with no reviews, workers often indicated they did not have enough information to make a recommendation judgement. Furthermore, it seems that systems that focus more reviews for ranking are relatively more effective for recommendation than for topical relevance. This suggests, not surprisingly, that assessors mainly base their recommendations on reviews.

Next year we will look more carefully at different aspects of relevance, such as topical relevance, recommendation, reading level and whether a books looks interesting or engaging. We also plan to analyse the suggestions in more detail and differentiate between books suggested by single and multiple forum members, positive and negative suggestions and suggested books that the topic creator decided to add to her personal catalogue.

Judging metadata or book content

In a realistic scenario, a user judges the relevance or interestingness of the book metadata, not of the content of the book. The decision to read a book comes before the judgement of the content. This points at an important problem with the suggested books collected through the touchstones. Members often suggest books they have actually read, and therefore base their suggestion on the actual content of the book. Such a relevance judgement—from someone other than the topic creator—is very different in nature from the judgement that the topic creator can make about books she has not read. Considering the suggested books as relevant brushes over this difference. We will further analyse the relevance and recommendation judgements from AMT to find out to what extent the LT forum suggestions reflect traditional topical relevance judgements and to what extent they reflect recommendation.

Extending the Collection
The Amazon/LibraryThing collection has a limited amount of professional metadata. Only 61% of the books have a DDC code and the Amazon subjects are noisy with many seemingly unrelated subject headings assigned to books. To make sure there is enough high-quality metadata from traditional library catalogues, we will extend the data set next year with library catalogue records from the Library of Congress and the British Library. We only use library records of ISBNs that are already in the collection. These records contain formal metadata such as classification codes (mainly DDC and LCC) and rich subject headings based on the Library of Congress Subject Headings (LCSH).[10] Both the LoC records and the BL records are in MARCXML[11] format. We obtained MARCXML records for 1.76 million books in the collection. Although there is no single library catalogue that covers all books available on Amazon, we think these combined library catalogues can improve both the quality and quantity of professional book metadata.

4 The Prove It (PI) Task

The goal of this task was to investigate the application of focused retrieval approaches to a collection of digitized books. The scenario underlying this task is that of a user searching for specific information in a library of books that can provide evidence to confirm or reject a given factual statement. Users are assumed to view the ranked list of book parts, moving from the top of the list down, examining each result. No browsing is considered (only the returned book parts are viewed by users).

Participants could submit up to 10 runs. Each run could contain, for each of the 83 topics (see Section 4.2), a maximum of 1,000 book pages estimated relevant to the given aspect, ordered by decreasing value of relevance.

A total of 18 runs were submitted by 2 groups (6 runs by UMass Amhers (ID=50) and 12 runs by Oslo University College (ID=100)), see Table 1.

4.1 The Digitized Book Corpus

The track builds on a collection of 50,239 out-of-copyright books[12], digitized by Microsoft. The corpus is made up of books of different genre, including history books, biographies, literary studies, religious texts and teachings, reference works, encyclopaedias, essays, proceedings, novels, and poetry. 50,099 of the books also come with an associated MAchine-Readable Cataloging (MARC) record, which contains publication (author, title, etc.) and classification information. Each book in the corpus is identified by a 16 character long bookID – the name of the directory that contains the book's OCR file, e.g., A1CD363253B0F403.

[10] For more information see: http://www.loc.gov/aba/cataloging/subject/
[11] MARCXML is an XML version of the well-known MARC format. See: http://www.loc.gov/standards/marcxml/
[12] Also available from the Internet Archive (although in a different XML format)

The OCR text of the books has been converted from the original DjVu format to an XML format referred to as BookML, developed by Microsoft Development Center Serbia. BookML provides additional structure information, including markup for table of contents entries. The basic XML structure of a typical book in BookML is a sequence of pages containing nested structures of regions, sections, lines, and words, most of them with associated coordinate information, defining the position of a bounding rectangle ([coords]):

```
<document>
 <page pageNumber="1" label="PT_CHAPTER" [coords] key="0" id="0">
  <region regionType="Text" [coords] key="0" id="0">
   <section label="SEC_BODY" key="408" id="0">
    <line [coords] key="0" id="0">
     <word [coords] key="0" id="0" val="Moby"/>
     <word [coords] key="1" id="1" val="Dick"/>
    </line>
    <line [...]><word [...] val="Melville"/>[...]</line>[...]
   </section>   [...]
  </region>     [...]
 </page>        [...]
</document>
```

BookML provides a set of labels (as attributes) indicating structure information in the full text of a book and additional marker elements for more complex structures, such as a table of contents. For example, the first label attribute in the XML extract above signals the start of a new chapter on page 1 (label="PT_CHAPTER"). Other semantic units include headers (SEC_HEADER), footers (SEC_FOOTER), back-of-book index (SEC_INDEX), table of contents (SEC_TOC). Marker elements provide detailed markup, e.g., for table of contents, indicating entry titles (TOC_TITLE), and page numbers (TOC_CH_PN), etc.

The full corpus, totaling around 400GB, was made available on USB HDDs. In addition, a reduced version (50GB, or 13GB compressed) was made available for download. The reduced version was generated by removing the word tags and propagating the values of the `val` attributes as text content into the parent (i.e., line) elements.

4.2 Topics

We use the same topic set as last year [6], consisting of 83 topics. Last year, relevance judgements were collected for 21 topics from two sources. In the first phase, INEX participants judged pages using the relevance assessment system developed at Microsoft Research Cambridge.[13] In the second phase, relevance judgements were collected from Amazon Mechanical Turk.

[13] URL: http://www.booksearch.org.uk

Table 13. Results for the 2011 Prove It evaluation using the 21 topics and judgements of the 2010 Prove It task. The run names of participant p100 (UMass) have been shortened to fit on the page. Best scores are in bold.

Run	MAP	MRR	P@10	nDCG@10 (0-1-2)	nDCG@10 (0-1-10)
p100-spec_10x_ge_55.res	0.0285	0.3271	0.1667	0.1238	0.0862
p100-spec_2x_ge_55.res	0.0255	0.3015	0.1619	0.1144	0.0788
p100-spec_5x_ge_55.res	0.0278	0.2995	0.1571	0.1145	0.0767
p100-to_g_10xover2.res	0.0491	0.4137	0.2810	0.2046	0.1469
p100-to_g_2xover2.res	0.0488	0.3855	0.2714	0.1963	0.1406
p100-to_g_5xover2.res	0.0490	0.4102	0.2762	0.2013	0.1439
p50.sdm.pass100.lambda0.025	0.3360	**1.0000**	0.7905	0.7809	0.7358
p50.sdm.pass50.lambda0.025	**0.3364**	**1.0000**	**0.8000**	**0.7842**	0.7365
p50.sdm	0.3330	**1.0000**	0.7905	0.7806	0.7356
p50.stopped.sdm.pass100.lambda0.025	0.3172	0.9762	0.7905	0.7767	0.7364
p50.stopped.sdm.pass50.lambda0.025	0.3177	0.9762	0.7905	0.7771	0.7369
p50.stopped.sdm	0.3136	0.9762	0.7905	0.7772	**0.7382**

4.3 Collected Relevance Assessments

The 2011 topic set is the same as the 2010 topic set, consisting of 21 topics. We reuse the judgements from last year and extend them with judgements based on top 10 pools of the official submissions. The UMass group (participant ID p100) provided their own judgements which we were kindly allowed to use for evaluation. In total, they judged 535 extra pages on top of the 2010 data set. The top 10 results of the runs submitted by OUC (participant ID p50) were pooled and judged using Mechanical Turk.

We used roughly the same design as last year [5], but with 6 pages per HIT instead of 10 and paying $0.30 per HIT, resulting in 92 HITs. Also, instead of asking workers to type the first word of the confirming/refuting sentence, we ask them to click on the first word of that sentence in the book page. We log the clicks, which allows us to check whether workers clicked inside the book page. This gives 419 new page-level judgements: 352 non-relevant pages, 21 relevant pages, 2 pages refuting the factual statement and 44 confirming it.

4.4 Evaluation Measures and Results

Similar to last year, the official evaluation measure is nDCG@10. Pages that confirm or refute a statement have a relevance value rv=2 and pages that are merely related to the topic have a relevance value rv=1. The evaluation results are shown in Table 13. As an alternative evaluation, we use judgements with extra weight on the confirm refute pages. The scores in column 6 in the table represent scores for the judgment where confirm/refute pages are weighted 10 (rv=10) times as much as pages that are merely relevant (rv=1). The runs submmited by UMass score very high on the official measure nDCG@10 and

three runs (starting with p50.sdm) get a perfect score on MRR. Their runs are based on Sequential Dependence Modelling, which is an interpolation between three language models based unigrams, bigrams and proximity respectively. By adjusting the Dirichlet smoothing parameter to the average number of words per page ($\mu = 363$), the SDM model is very effective in locating confirming and refuting pages.

The evaluation results of this year show that the current Prove It task can be adequately solved. For next year's Prove It task, we will introduce further challenges in identifying confirming and refuting information. One possibility is to use the confirm/refute label in the evaluation measure. That is, systems have to determine whether a page confirms or refutes a statement. As most systems find almost no refuting pages, it would seem that a trivial solution of labelling all returned results as confirming would score very high. This could be used as a baseline, which might encourage participants to focus more on finding refute pages so as to beat this baseline.

Again, the complexity of the factual statement of many topics caused problems for assessors. The statements often consist of multiple atomic facts, which confronts assessors with the problem of deciding whether a page confirms or refutes a statement when only one or some of the atomic facts are confirmed or refuted. A possible solution may be to make the topics more structured by splitting complex statements into their atomic parts, and asking assessors to judge pages on each part of the statement.

5 The Structure Extraction (SE) Task

The goal of the SE task was to test and compare automatic techniques for extracting structure information from digitized books and building a hyperlinked table of contents (ToC). The task was motivated by the limitations of current digitization and OCR technologies that produce the full text of digitized books with only minimal structure markup: pages and paragraphs are usually identified, but more sophisticated structures, such as chapters, sections, etc., are typically not recognised.

In 2011, the task was run for the second time as a competition of the International Conference on Document Analysis and Recognition (ICDAR). Full details are presented in the corresponding specific competition description [4]. This year, the main novelty was the fact that the ground truth data built in 2009 and 2010 was made available online [14]. Participants were hence able to build and fine tune their systems using training data.

Participation

Following the call for participation issued in January 2011, 11 organizations registered. As in previous competitions, several participants expressed interest

[14] http://users.info.unicaen.fr/~doucet/StructureExtraction/training/

but renounced due to time constraints. Of the 11 organizations that signed up, 5 dropped out; that is, they neither submitted runs, nor participated in the ground truth annotation process. The list of active participants is given in Table 14. Interestingly, half of them are newcomers (Nankai University, NII Tokyo and University of Innsbrück).

Organization	Submitted runs	Ground truthing
Microsoft Development Center (Serbia)	1	y
Nankai University (PRC)	4	y
NII Tokyo (Japan)	0	y
University of Caen (France)	3	y
University of Innsbrück (Austria)	0	y
Xerox Research Centre Europe (France)	2	y

Table 14. Active participants of the Structure Extraction task.

Results

As in previous years [3], the 2011 task permitted to gather manual annotations in a collaborative fashion. The efforts of the 2011 round gave way to the gathering and addition of 513 new annotated book ToCs to the previous 527.

A summary of the performance of all the submitted runs is given in Table 15.

RunID	Participant	Title-based [3]	Link-based [2]
MDCS	MDCS	40.75%	65.1%
Nankai-run1	Nankai U.	33.06%	63.2%
Nankai-run4	Nankai U.	33.06%	63.2%
Nankai-run2	Nankai U.	32.46%	59.8%
Nankai-run3	Nankai U.	32.43%	59.8%
XRCE-run1	XRCE	20.38%	57.6%
XRCE-run2	XRCE	18.07%	58.1%
GREYC-run2	University of Caen	8.99%	50.7%
GREYC-run1	University of Caen	8.03%	50.7%
GREYC-run3	University of Caen	3.30%	24.4%

Table 15. Summary of performance scores for the Structure Extraction competition 2011 (F-measures).

The Structure Extraction task was launched in 2008 to compare automatic techniques for extracting structure information from digitized books. While the

construction of hyperlinked ToCs was originally thought to be a first step on the way to the structuring of digitized books, it turns out to be a much tougher nut to crack than initially expected.

Future work aims to investigate into the usability of the extracted ToCs. In particular we wish to use qualitative measures in addition to the current precision/recall evaluation. The vast effort that this requires suggests that this can hardly be done without crowdsourcing. We shall naturally do this by building on the experience of the Book Search tasks described earlier in this paper.

6 The Active Reading Task (ART)

The main aim of the Active Reading Task (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. The task is motivated by the lack of common practices when it comes to conducting usability studies of e-reader tools. Current user studies focus on specific content and user groups and follow a variety of different procedures that make comparison, reflection, and better understanding of related problems difficult. ART is hoped to turn into an ideal arena for researchers involved in such efforts with the crucial opportunity to access a large selection of titles, representing different genres, as well as benefiting from established methodology and guidelines for organising effective evaluation experiments.

The ART is based on the evaluation experience of EBONI [8], and adopts its evaluation framework with the aim to guide participants in organising and running user studies whose results could then be compared.

The task is to run one or more user studies in order to test the usability of established products (e.g., Amazon's Kindle, iRex's Ilaid Reader and Sony's Readers models 550 and 700) or novel e-readers by following the provided EBONI-based procedure and focusing on INEX content. Participants may then gather and analyse results according to the EBONI approach and submit these for overall comparison and evaluation. The evaluation is task-oriented in nature. Participants are able to tailor their own evaluation experiments, inside the EBONI framework, according to resources available to them. In order to gather user feedback, participants can choose from a variety of methods, from low-effort online questionnaires to more time consuming one to one interviews, and think aloud sessions.

6.1 Task Setup

Participation requires access to one or more software/hardware e-readers (already on the market or in prototype version) that can be fed with a subset of the INEX book corpus (maximum 100 books), selected based on participants' needs and objectives. Participants are asked to involve a minimum sample of

15/20 users to complete 3-5 growing complexity tasks and fill in a customised version of the EBONI subjective questionnaire, allowing to gather meaningful and comparable evidence. Additional user tasks and different methods for gathering feedback (e.g., video capture) may be added optionally. A crib sheet is provided to participants as a tool to define the user tasks to evaluate, providing a narrative describing the scenario(s) of use for the books in context, including factors affecting user performance, e.g., motivation, type of content, styles of reading, accessibility, location and personal preferences.

Our aim is to run a comparable but individualized set of studies, all contributing to elicit user and usability issues related to eBooks and e-reading.

7 Conclusions and plans

This paper presents an overview of the INEX 2011 Books and Social Search Track. The track has four tasks: 1) Social Search for Best Books, 2) Prove It, 3) Structure Extraction, and 4) Active Reading Task.

This was the first year for The Social Search for Best Books (SB) task, but the amount of activity and the results promise a bright future for this task. The comparison of the LT forum suggestions and the relevance judgements from the Mechanical Turk experiment show that forum suggestions represent a different task from traditional ad hoc topical relevance search. The two sets of judgements give us an interesting data set to address questions about the relative value of professional controlled metadata and user-generated content for book search, for subject search topics as well as more recommendation oriented topics.

Preliminary analysis of the crowdsourcing data suggests that assessors treat topical relevance and recommendation similarly. If they consider book topically relevant, they often recommend it and vice versa, do not recommend when it is not on the right topic. Next year, we want to focus more specifically on the various aspects of relevance for book suggestions, such as topical relevance, recommendation, reading level, engagement etc.

This year the Prove It task continued unchanged with respect to last year. The number of participants for the PI task was low. We gathered relevance judgements from participants and from Mechanical Turk based on top 10 pools. The Mechanical Turk experiment is still running, but preliminary results show that the runs submitted by the University of Massachusetts Amherst leave little room for improvement. We will introduce new interesting challenges in the Prove It task for next year. One idea is to require systems to indicate whether a page contains confirming or refuting information.

In 2011, the SE track was run conjointly within the ICDAR conference for the second time. This effort gave way to the gathering and addition of 513 new annotated book ToCs to the previous 527, available for download on the track's Web site. The SE task will be run again at ICDAR 2013.

The ART was offered as last year. The task has so far only attracted 2 groups, none of whom submitted any results at the time of writing.

Bibliography

[1] Thomas Beckers, Norbert Fuhr, Nils Pharo, Ragnar Nordlie, and Khairun Nisa Fachry. Overview and results of the inex 2009 interactive track. In Mounia Lalmas, Joemon M. Jose, Andreas Rauber, Fabrizio Sebastiani, and Ingo Frommholz, editors, *ECDL*, volume 6273 of *Lecture Notes in Computer Science*, pages 409–412. Springer, 2010.

[2] Hervé Déjean and Jean-Luc Meunier. Reflections on the inex structure extraction competition. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, pages 301–308, New York, NY, USA, 2010. ACM.

[3] Antoine Doucet, Gabriella Kazai, Bodin Dresevic, Aleksandar Uzelac, B.Radakovic, and Nikola Todic. Setting up a competition framework for the evaluation of structure extraction from ocr-ed books. *International Journal of Document Analysis and Recognition (IJDAR), Special Issue on Performance Evaluation of Document Analysis and Recognition Algorithms.*, 14(1):45–52, 2011.

[4] Antoine Doucet, Gabriella Kazai, and Jean-Luc Meunier. ICDAR 2011 Book Structure Extraction Competition. In *Proceedings of the Eleventh International Conference on Document Analysis and Recognition (ICDAR'2011)*, pages 1501–1505, Beijing, China, September 2011.

[5] Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 205–214. ACM Press, New York NY, 2011.

[6] Gabriella Kazai, Marijn Koolen, Jaap Kamps, Antoine Doucet, and Monica Landoni. Overview of the INEX 2010 book track: Scaling up the evaluation using crowdsourcing. In Shlomo Geva, Jaap Kamps, Ralf Schenkel, and Andrew Trotman, editors, *Comparative Evaluation of Focused Retrieval : 9th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2010)*, volume 6932 of *LNCS*, pages 101–120. Springer, 2011.

[7] Elaine Svenonius. *The Intellectual Foundation of Information Organization.* MIT Press, 2000.

[8] Ruth Wilson, Monica Landoni, and Forbes Gibb. The web experiments in electronic textbook design. *Journal of Documentation*, 59(4):454–477, 2003.