Overview of the INEX 2013 Social Book Search Track

Koolen, M.; Kazai, G.; Preminger, M.; Doucet, A.

Publication date
2013
Document Version
Final published version
Published in
CLEF 2013 : CLEF2013 Working Notes

Overview of the INEX 2013 Social Book Search Track

Marijn Koolen[1], Gabriella Kazai[2], Michael Preminger[3], and Antoine Doucet[4]

[1] University of Amsterdam, Netherlands
{marijn.koolen,kamps}@uva.nl
[2] Microsoft Research, United Kingdom
a-gabkaz@microsoft.com
[3] Oslo and Akershus University College of Applied Sciences, Norway
michaelp@hioa.no
[4] University of Caen, France
doucet@info.unicaen.fr

Abstract. The goal of the INEX 2013 Social Book Search Track is to evaluate approaches for supporting users in reading, searching, and navigating collections of books based on book metadata, the full texts of digitised books or associated user-generated content. The investigation is focused around three tasks: 1) the Social Book Search (SBS) task investigates the complex nature of relevance in book search and the role of traditional and user-generated book metadata in retrieval, 2) the Prove It (PI) task evaluates focused retrieval approaches for searching pages in books that can confirm or refute a given factual claim, 3) the Structure Extraction (SE) task evaluates automatic techniques for deriving book structure from OCR text and layout information. Both the SBS and SE tasks have a growing number of active participants, while the PI task is only tackled by a small number of core groups. In the SBS task, we extended last year's investigation into the nature of book suggestions from the LibraryThing forums and how they compare to book relevance judgements. We found further support that such suggestions are a valuable alternative to traditional test collections that are based on top-k pooling and editorial relevance judgements. The PI task added a further relevance criterion that pages should not only confirm or refute a given factual claim, but should also come from an authoritative source that is of the appropriate genre. The relevance assessments have not yet commenced at the time of writing. The SE task has reached a record number of active participants and has, for the first time, witnessed an improvement in the state of the art.

1 Introduction

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 digitised 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 four key areas:

- Evaluation methodologies for book search tasks that combine aspects of retrieval and recommendation,
- Information retrieval techniques for dealing with professional and user-generated metadata,
- Semantic and focused retrieval techniques for searching collections of digitised books, and
- Mechanisms to increase accessibility to the contents of digitised books.

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

1. 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.
2. The *Prove It* (PI) task aims to evaluate focused retrieval approaches on collections of digitised books, where users expect to be pointed directly at relevant pages that may help to confirm or refute a given factual claim;
3. The *Structure Extraction* (SE) task aims at evaluating automatic techniques for deriving structure from OCR and building hyperlinked table of contents.

In this paper, we report on the setup and the results of each of these tasks at the 2013 INEX@CLEF Lab. First, in Section 2, we give a brief summary of the participating organisations. The SBS task is described in detail in Section 3, the PI task in Section 4, and the SE task in Section 5. We close in Section 6 with a summary and plans for 2014.

2 Participating Organisations

A total of 68 organisations registered for the track (compared with 55 in 2012 and 47 in 2011). At the time of writing, we counted 14 active groups (compared with 5 in 2012 and 10 in 2011), see Table 1.

3 The Social Book Search Task

The goal of the Social Book Search (SBS) task is to evaluate the value of professional metadata and user-generated content for book search on the Web and

Table 1. Active participants of the INEX 2013 Social Book Search Track, the task(s) they were active in, and number of contributed runs (SBS = Social Book Search, PI = Prove It , SE = Structure Extraction)

ID	Institute	Tasks	Runs
4	University of Amsterdam, ILLC	SBS, PI	6 SBS, 5 PI
4	University of Amsterdam, ILPS	SBS	2 SBS
54	Royal School of Library and Information Science	SBS	3 SBS
100	Oslo & Akershus University College of Applied Sciences	SBS, PI	3 SBS, 6 PI
113	University of Caen	SE	5 SE
123	LSIS / Aix-Marseille University	SBS	3 SBS
147	National Taiwan Normal University	SBS	6 SBS
180	Chaoyang University of Technology	SBS	6 SBS
232	Indian School of Mines, Dhanbad	SBS	5 SBS
280	University of Würzburg	SE	1 SE
288	University of Innsbruck	SE	1 SE
299	EPITA/LRDE	SE	1 SE
303	Nankai University	SE	1 SE
	Microsoft Development Center Serbia	SE	1 SE
	Total		34 SBS, 11 PI, 10 SE

to develop and evaluate systems that can deal with both retrieval and recommendation aspects, where the user has a specific information need against a background of personal tastes, interests and previously seen books. Through social media, book descriptions have extended far beyond what is traditionally stored in professional catalogues. Not only are books described in the users' own vocabulary, but are also reviewed and discussed online, and added to online personal catalogues of individual readers. This additional information is subjective and personal, and opens up opportunities to aid users in searching for books in different ways that go beyond the traditional editorial metadata based search scenarios, such as known-item and subject search. For example, readers use many more aspects of books to help them decide which book to read next [7], such as how engaging, fun, educational or well-written a book is. In addition, readers leave a trail of rich information about themselves in the form online profiles which contain personal catalogues of the books they have read or want to read, personally assigned tags and ratings for those books and social network connections to other readers. This results in a search task that may require a different model than traditional ad hoc search [5] or recommendation.

The SBS task investigates book requests and suggestions from the LibraryThing (LT) discussion forums as a way to model book search in a social environment. The discussions in these forums show that readers frequently turn to others to get recommendations and tap into the collective knowledge of a group of readers interested in the same topic.

The task builds on the INEX Amazon/LibraryThing (A/LT) collection [1], which contains 2.8 million book descriptions from Amazon, enriched with content from LT. This collection contains both professional metadata and user-generated content.

The SBS task aims to address the following research questions:

- Can we build reliable and reusable test collections for social book search based on book requests and suggestions from the LT discussion forums?
- Can user profiles provide a good source of information to capture personal, affective aspects of book search information needs?
- How can systems incorporate both specific information needs and general user profiles to combine the retrieval and recommendation aspects of social book search?
- What is the relative value of social and controlled book metadata for book search?

3.1 Scenario

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

On LT, 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 discussion forums 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 known books with the intention to obtain them (buy, download, print). Such needs are addressed by standard book search services as offered by Amazon, LT and other online bookshops as well as traditional libraries. In other cases, users search for a specific, but unknown, book with the intention of identifying it. Another possibility is that users 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 or that are similar to a previously read book.

3.2 Task description

The SBS task is to reply to a user request posted on a LT forum (see Section 3.5) by returning a list of recommended books matching the user's information need. More specifically, the task assumes a user who issues a query to a retrieval system, which then returns a (ranked) list of relevant book records. The user is assumed to inspect the results list starting from the top, working down the

list until the information need has been satisfied or until the user gives up. The retrieval system is expected to order the search results by relevance to the user's information need.

The user's query can be a number of keywords, but also one or more book records as positive or negative examples. In addition, the user has a personal profile that may contain information on the user's interests, list of read books and connections with other readers. User requests may vary from asking for books on a particular genre, looking for books on a particular topic or period or books written in a certain style. 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.

The books must be selected from a corpus that consists of a collection of curated and social book metadata, extracted from Amazon Books and LT, extended with associated records from library catalogues of the Library of Congress and the British Library (see the next section). Participants of the SBS task are provided with a set of book search requests and user profiles and are asked to submit the results returned by their systems as ranked lists.

The SBS task, thus, combines aspects from retrieval and recommendation. On the one hand, the task is akin to directed search, familiar from information retrieval, with the requirement that returned books should be topically relevant to the user's information need described in the forum thread. On the other hand, users may have particular preferences for writing style, reading level, knowledge level, novelty, unusualness, presence of humorous elements and possibly many other aspects. These preferences are to some extent reflected by the user's reading profile represented by the user's personal catalogue. This catalogue contains the books already read or earmarked for future reading, and may contain personally assigned tags and ratings. Such preferences and profiles are typical in recommendation tasks, where the user has no specific information need, but is looking for suggestions of new items based on previous preferences and history.

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 LT 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 identify 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 (A/LT) corpus crawled by the University of Duisburg-Essen for the INEX Interactive Track [1]. See https://inex.mmci.uni-saarland.de/data/nd-agreements.jsp for information on how to gain access to the corpus.

The collection consists of 2.8 million book records from Amazon, extended with social metadata from LT. This set represents the books available through Amazon. The records contain title information as well as a Dewey Decimal Classification (DDC) code (for 61% of the books) and category and subject information supplied by Amazon. We note that for a sample of Amazon records we noticed the subject descriptors to be noisy, with a number of inappropriately assigned descriptors that seem unrelated to the books.

Each book is identified by an ISBN. Since different editions of the same work have different ISBNs, there can be multiple records for a single intellectual work. Each book record is an XML file with fields like *isbn*, *title*, *author*, *publisher*, *dimensions*, *numberofpages* and *publicationdate*. Curated metadata comes in the form of a Dewey Decimal Classification in the *dewey* field, Amazon subject headings in the *subject* field, and Amazon category labels in the *browseNode* fields. The social metadata from Amazon and LT is stored in the *tag*, *rating*, and *review* fields. The full list of fields is shown in Table 2.

To ensure that there is enough high-quality metadata from traditional library catalogues, we extended the A/LT data set with library catalogue records from the Library of Congress (LoC) and the British Library (BL). We only use library records of ISBNs that are already in the A/LT collection. These records contain formal metadata such as title information (book title, author, publisher, etc.), classification codes (mainly DDC and LCC) and rich subject headings based on the Library of Congress Subject Headings (LCSH).[5] Both the LoC records and the BL records are in MARCXML[6] format. There are 1,248,816 records from the LoC and 1,158,070 records in MARC format from the BL. Combined, there are 2,406,886 records covering 1,823,998 of the ISBNs in the A/LT collection (66%). Although there is no single library catalogue that covers all books available on

[5] For more information see: http://www.loc.gov/aba/cataloging/subject/

[6] MARCXML is an XML version of the well-known MARC format. See: http://www.loc.gov/standards/marcxml/

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	

Amazon, we reason that these combined library catalogues can improve both the quality and quantity of professional book metadata. Indeed, with the LoC and BL data sets combined, 79% of all ISBNs in the original A/LT corpus now have a DDC code. In addition, the LoC data set also has LCC codes for 44% of the records in the collection. With only the A/LT data, 57% of the book descriptions have at least one subject heading, but with the BL and LoC data added, this increases to 80%. Furthermore, the A/LT data often has only a single subject heading per book, whereas in the BL and LoC data sets, book descriptions typically have 2–4 headings (average 2.96). Thus, the BL and LoC data sets increase the coverage of curated metadata, such that the vast majority of descriptions in our data set include professionally assigned classification codes and subject headings.

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 of the same work 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 ignore these distinctions to make evaluation easier. This turns edition-related topics into known-item topics.

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.[7]

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 any time, some ISBNs map to multiple work IDs even though they represent the same intellectual work. 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.

3.5 Information needs

LT users discuss their books on the discussion forums. Many of the topic threads are started with a request from a member for interesting, fun new books to read. Users typically 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. Members often reply with links to works catalogued on LT, which have direct links to the corresponding records on Amazon. These requests for recommendations are natural expressions of information needs for a large collection of online book records. We use a sample of these forum topics to evaluate systems participating in the SBS task.

Each topic has a title and is associated with a group on the discussion forums. For instance, topic 99309 in Figure 1 has the title *Politics of Multiculturalism Recommendations?* and was posted in the group *Political Philosophy*. The books suggested by members in the thread are collected in a list on the side of the topic thread (see Figure 1). A feature 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 in LT, with associated ISBNs and links to Amazon. We use these suggested books as initial relevance judgements for evaluation. In the rest of this paper, we use the term *suggestion* to refer to a book that has been identified in a touchstone list for a given forum topic. Since all suggestions are made by forum members, we assume they are

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

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

valuable judgements on the relevance of books. Additional relevance information can be gleaned from the discussions on the threads. Consider, for example, topic 129939[8]. The topic starter first explains what sort of books he is looking for, and which relevant books he has already read or is reading. Other members post responses with book suggestions. The topic starter posts a reply describing which suggestions he likes and which books he has ordered and plans to read. Later on, after some more discussions, the topic starter provides feedback on the suggested books that he has now read. Such feedback can be used to estimate the relevance of a suggestion to the user.

User profiles and personal catalogues From LT we can not only extract the information needs of social book search topics, but also the rich user profiles of the topic starters, 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, transformed into XML and made available to participants. This adds a recommendation aspect to the task. One of the research questions of the SBS task is whether this profile information can help systems in identifying good suggestions.

Although the user expresses her information need in some detail in the discussion forum, she may not describe all aspects she takes into consideration

[8] URL: urlhttp://www.librarything.com/topic/129939

Table 3. User profile statistics of the topic starters. Days of membership is the number of days between becoming LT member and posting the topic on the forum

Type	N	min	max	median	mean	stdev
Days of membership	373	0	2093	215	413.08	476.50
Friends	380	0	135	2	7.21	15.28
Interesting libraries	380	0	266	0	10.33	26.20
Groups	380	0	10	6	5.64	4.18
Books	380	0	6294	104	505.78	1002.35
Ratings	380	0	2771	15	163.59	389.85
Tags	380	0	44,283	191	1549.19	4160.44

when selecting books. This may partly be because she wants to explore different options along different dimensions and therefore leaves some room for different interpretations of her need. Another reason might be that some aspects are not related directly to the topic at hand but may be latent factors that she takes into account with selecting books in general.

What information is available in the LT profile of the topic starter? First, there are the connections to other members, which come in two kinds, *friendship* connections and *interesting library* connections. The former is a connection with another LT member she knows, the latter a connection to an LT member whose personal catalogue she finds interesting. Then there are the *group* membership connections to the discussion groups on the LT forums. These signal some of her (book-related) interests. Finally, there is the personal catalogue, which contains basic information on the books that the user has added to her personal LT catalogue, when she added each book, as well as any tags and ratings she assigned to each individual book. Basic statistics on the number of connections, books, ratings and tags per user profile is given in Table 3. In the top row we see the number of days between the topic starter registering to LT and posting the topic on the forum. On average, users build up their profiles for over a year before they ask for book suggestions, although the distribution is skewed. Most book recommendation requests are posted within 215 days of registering to LT, with 54 topics (14%) posted on the day of registration. The latter topics may actually be the reason that the topic starter registered to LT. All frequency distributions, with the exception of group connections, are heavily skewed, with a small number of high end outliers causing the mean to be much higher than the median. The number of groups shown on the profile is cut-off by LT at 10 groups, which means our profiles miss some group connections for members with more than 10 groups. Most topic starters have a small number of friends and are registered to up to half a dozen discussion groups. Topic starters have around 100 books in their catalogues, almost twice as many tags and only few ratings, although some users have much bigger and more heavily tagged catalogues. There is a lot of variation in the profiles in all aspects mentioned in Table 3. Especially

for profiles with many books, tags and connections, recommendation algorithms may need to be tuned to the parts of the profile that are relevant to the specific information need of the topic thread.

In the following sections, we describe the procedures for topic selection (Section 3.6) and for suggestion annotation (Section 3.7) procedure, then how we used the annotations to assign relevance values to the suggestions (Section 3.8).

3.6 Topic selection

Three students from the Royal School of Library and Information Science were paid to annotate the narratives of a sample of LT forum topics. We created an interface to help them to 1) select topic threads that are about book search information needs (as opposed to anything else that people discuss on the LT forums), 2) annotate the selected topics describing the type of information need—are users looking for books about a particular topic, in a certain genre, by a certain author, etc.—and 3) annotate the suggestions provided by LT members in the thread. The latter included questions on whether the suggestors have read the suggested books and what their attitudes were towards the books, i.e. positive recommendations vs negative mentions.

An initial set of 9,401 topic threads containing touchstones was made available to the students. Of those, the students annotated over 1100 topics, of which 386 (35%) were identified as topics with a book search information need. From the 386, six topics contain no suggestions to any of the books in our A/LT collection. Although all 386 topics were distributed to participants, these 6 topics are discarded from the evaluation.

Topics can have complex information needs, with a combination of multiple relevance aspects. Traditionally, in IR, the focus has been on what a document is about, but in book search there are often many other aspects of relevance. Reuter [7] identified 7 general categories of relevance aspects for book search, to which we added the category of known-item information needs:

Metadata books with a certain title or by a certain author, editor, illustrator, publisher, in a particular format, or written

Accessibility the language, length or level of difficulty of a book,

Content aspects such as topic, plot, genre, style or comprehensiveness of a book,

Engagement books that fit a particular mood or interest, or books that are considered high quality or provide a particular reading experience,

Novelty books with novel content for the reader, books that are unusual,

Familiarity similar to known book or related to previous experience,

Socio-Cultural books related the user's socio-cultural background or values, books that are popular or obscure.

Known-item description of known book to identify title and/or author, or published in certain year or period,

In the second step, annotators had to indicate which aspects of relevance the topics relate to. Annotators could select multiple relevance categories. For example, for topic 99309 on the *politics of muliticulturalism*, the topic starter asks for

Table 4. Distribution of relevance aspect types over topics

Aspect	#	%
Metadata	138	36
Accessibility	63	17
Content	314	83
Engagement	107	28
Novelty	13	3
Familiarity	168	44
Socio-Cultural	58	15
Known-item	92	24

Table 5. Distribution of number of relevance aspects per topic

# Aspects	# topics	(%)
1	57	(15%)
2	144	(38%)
3	134	(35%)
4	37	(10%)
5	6	(2%)
6	2	(1%)

suggestions about a particular topic—i.e., *content* relevance—but also asks for books that add something new to what he has already read on the topic—i.e., *novelty*. The distribution of the relevance aspects in the topic set is shown in Table 4. Book search information needs on the LT forums almost always (83% of the 380 topics) contain content aspect. This reinforces the traditional choice in designing best match retrieval models around aspects of document content. Metadata aspects, such as book title and author, are present in 36% of the data set. Other important aspects are *familiarity* (44%) and *engagement* (28%) and *known-item* (24%). Looking for books similar to certain books a user has read is the task of item-based recommender systems, such as that offered by Amazon ('*customers who bought this book also bought...*'). It reinforces our interpretation of LT forum book search needs as a task that combines aspects of retrieval and recommendation. Engagement is something that is hard to express in a search engine query. For instance, how can a user search for text books that are funny or high-brow literature that is scary, or books that challenge the reader's own views on a topic? So it is not surprising that readers instead turn to other readers to ask for suggestions. The same holds for read or 'heard about' books for which the user only recalls some aspect of the plot, or the some attributes of certain characters. Book search services are of limited use for such known-item topics, but forum members might be able to help out. *Accessibility*, *novelty* and *socio-cultural* aspects are less prominent in our sample set. Only 15% of the topics have a single relevance aspect (see Table 5). The vast majority topics represent

complex information needs—most topics have 2 or 3 relevance aspects, 144 and 134 topics respectively—with a mean number of 2.47 aspects per topic.

In addition to the above, annotators had to indicate whether the request was for fiction, non-fiction or both and they had to provide a search query that they would use with a book search engine. We hope that these queries better express the information need than some of the topic thread titles, some of which do not describe the information need at all. Of the 380 topics, 175 (46%) asked for suggestions on fiction books, 49 (13%) on non-fiction, 55 (14%) on both fiction and non-fiction, and for 101 topics (27%) the annotator could not tell. The fraction of non-fiction topics is lower than last year (49%). We assume that this difference is caused by giving the annotators the option to indicate they were *not sure*, whereas in last year's topic selection procedure, we always choose between fiction, non-fiction or both.

Figure 1 shows an annotated topic (topic 99309) as an example:

```
<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>
</topic>
```

3.7 Suggestion annotations

Finally, annotators had to label each book suggestion provided by LT members (including any provided by the topic starter). They had to indicate whether the suggestor *has read* the book. For the *has read* question, the possible answers were *Yes*, *No*, *Can't tell* and *It seems like this is not a book*. They also had to judge the attitude of the suggestor towards the book. Possible answers were *Positively*, *Neutrally*, *Negatively*, *Not sure* or *This book is not mentioned as a relevant suggestion!* The latter can be chosen when someone mentions a book for another reason than to suggest it as a relevant book for the topic of request.

The 380 topic threads were started by 378 different members in 347 different discussion groups, containing altogether 2,568 messages from 1,109 different LT

Table 6. Distribution of participants, messages and suggestions per thread, number of topics in which participants participate and the number of suggestions per message

description	#	min.	max.	med.	mean	std.dev
participants/topics	380	1	53	3	5.39	5.75
messages/topic	380	1	107	4	6.76	8.73
touchstones/topic	380	1	234	8	15.87	22.73
topics/participant	1109	1	68	2	4.26	6.06
touchstones/message	2568	1	204	10	17.55	25.83
touchstones/message	2568	1	29	1	2.35	2.55

Table 7. Distribution of annotation labels over the answer categories on whether the suggestor *has read* the suggestion and the suggestors *attitude* towards to book

	Attitude						
Read	pos	neu	neg	not sure	non sug	unticked	total
yes	2757	700	134	116	87	0	3794
no	198	263	15	11	15	0	502
can't tell	136	1224	7	97	107	0	1571
non book	2	0	0	0	0	160	162
total	3093	2187	156	224	209	160	6029

members. The distribution of participants, messages and touchstones per topic as shown in Table 6. On average, each thread has 5.39 participants (median is 3), 6.76 messages (median 4) and 15.87 touchstones (median 8). The distributions are all skewed to the right. Most participants contribute touchstones to multiple topics (topics/participant). Messages with touchstones typically contain only one suggestion (median touchstones per message is 1), although some messages contain many touchstones (58 messages contain 10 touchstones or more).

The relationship between having read the book and the suggestor's attitude towards the book is shown in Table 7. In the majority of cases (63%) members suggest books that they have read. It is rather rare for suggestors to state that they have not read a suggested book (8%). More often, suggestors do not reveal whether they have read the book or not (26%). Books mentioned in response to a book search request are often presented in a positive (51%) or neutral (36%) way. Both positive and negative suggestions tend to come from members who have read the books. When books are mentioned in a neutral way, it is often difficult to tell whether the book has been read by the suggestor, although a third of the neutral mentions comes from members who have read the book. There are 162 touchstones that do not refer to books but to author names. In almost all such cases (160 out of 162), annotators skipped the question regarding the attitude of the suggestor towards the book. In the two remaining cases, the annotator indicated the suggestor mentioned the author in a positive way.

How do the attitude labels relate to the books mentioned by the forum members? Are all books mentioned only once or is there a lot of discussion in many threads? Do members agree on which books are good suggestions and which ones are not? Within a thread, books can be mentioned multiple times, by a single member or by multiple members. The 6029 touchstones represent 5092 books. Most books, 4480 or 88% were mentioned only once in the thread, 612 (12%) were mentioned twice or more and 175 (3%) were mentioned three times or more. Of the 4480 books mentioned only once, 2292 (51%) are mentioned positively, 1683 (38%) neutrally, 85 (2%) negatively and 420 (9%) are labelled as either *not sure*, *non-suggestion* or were skipped because they were author names misidentified as book titles. Books mentioned by forum members tend to be positive suggestions. Of the 612 books mentioned multiple times, attitudes are mostly all positive (230 books, or 38%)) or a mix of positive and neutral (194 or 32%). For 116 books (19%), all attitudes are neutral. Only for 31 books (5%) there is real disagreement—some are positive, some are negative—and for 26 books (4%) attitudes are all negative or mixed neutral and negative. For 15 books (2%) the annotators cannot tell the attitude or indicated the touchstone is not actually a book. In other words, when books are mentioned multiple times, forum members rarely disagree with each other and are mostly positive.

All in all, in response to a book search request, members suggest mostly books they have read and often in a positive way. This supports our choice of using forum suggestions as relevance judgements.

3.8 Operationalisation of forum judgement labels

The annotated suggestions labels were used to determine the relevance value of each book suggestion in the thread. Because some of the books mentioned in the forums are not part of the 2.8 million books in our collection, we first removed from the suggestions any books that are not in the INEX A/LT collection, which leaves 4572 out of 5092 books (90%). For these 4572 books we derive relevance values.

Forum members can mention books for many different reasons. We want the relevance values to distinguish between books that were mentioned as positive recommendations, negative recommendations (books to avoid), neutral suggestions (mentioned as possibly relevant but not necessarily recommended) and books mentioned for some other reason (not relevant at all). We also want to differentiate between recommendations from members who have read the book they recommend and members who have not. We assume the recommendation to be of more value to the searcher if it comes from someone who has actually read the book. For the mapping to relevance values, we refer to the first mention of work as the *suggestion* and subsequent mentions of the same work as *replies*. We use *has read* when the forum members have read the book they mention and *not read* when they have not. Furthermore, we use a number of simplifying assumptions:

- When the annotator was *not sure* if the person mentioning a book has read it, we treat it as *not read*. We argue that for the topic starter it is not clear there is a difference in value of such recommendations.
- When the annotator was *not sure* if the person mentioning a book is positive, negative or neutral, we treat it as *neutral*. Again, for the topic starter there is no clear signal that there is difference in value.
- *has read* recommendations overrule *not read* recommendations. Someone who has read the book is in a better position to judge a book than of someone who has not.
- *positive* and *negative* recommendations neutralise each other. I.e. a *positive* and a *negative* recommendation together are the same as two *neutral* recommendations.
- If the topic starter *has read* a book she mentions, the relevance value is $rv = 0$. We assume such books have no value as suggestions.
- The attitude of the topic starter towards a book overrules those of others. The system should retrieve books for the topic starter, not for others.
- When forum members mention a single work multiple times, we use the last mention as judgement.

With the following decision tree we determine from which forum members want to use the judgements to derive relevance values:

1. Book mentioned by single member → use that member's judgement
2. Book mentioned by multiple members
   - 2.1 topic starter mentions book
     - 2.1.1 topic starter only suggests neutrally → use replies of others (2.2)
     - 2.1.1 topic starter suggests positively/negatively → use starter judgement
     - 2.1.1 topic starter replies → use starter judgement
   - 2.2 topic starter does not mention book
     - 2.2.2 members who have read the book suggest/reply → use *has read* judgements
     - 2.2.2 no member who suggests/replies about a book has read it → use all judgements

Once the judgements per suggested book are determined, we map the annotated judgements to relevance values. The base relevance value of a book that is mentioned in the thread is $rv = 2$. The values are modified according to the following scheme:

1. single judgement
   - 1.1 starter has read judgement → $rv = 0$
   - 1.2 starter has not read judgement
     - 1.2.2 starter positive → $rv = 8$
     - 1.2.2 starter neutral → $rv = 2$
     - 1.2.2 starter negative → $rv = 0$
   - 1.3 other member has read judgement
     - 1.3.3 has read positive → $rv = 4$

1.3.3 has read neutral → $rv = 2$
1.3.3 has read negative → $rv = 0$
1.4 other member has not read judgement
1.4.4 not read positive → $rv = 3$
1.4.4 not read neutral → $rv = 2$
1.4.4 not read negative → $rv = 0$
2. multiple judgements
2.1 multiple has read judgements
2.1.1 some positive, no negative → $rv = 6$
2.1.1 #positive > #negative → $rv = 4$
2.1.1 #positive == #negative → $rv = 2$
2.1.1 all neutral → rv=2
2.1.1 #positive < #negative → $rv = 1$
2.1.1 no positive, some negative → $rv = 0$
2.2 multiple not read judgements
2.2.2 some positive, no negative → $rv = 4$
2.2.2 #positive > #negative → $rv = 3$
2.2.2 #positive == #negative → $rv = 2$
2.2.2 all neutral → rv=2
2.2.2 #positive < #negative → $rv = 1$
2.2.2 no positive, some negative → $rv = 0$

This results in graded relevance values with seven possible values (0, 1, 2, 3, 4, 6, 8). For the 380 topics, there are 4572 relevance values (see Table 8), with 438 from judgements by the topic starter, and 4134 from the judgements of other members. Of these, 3892 are based on single judgements and 242 on judgements from multiple other members. The topic starters contribute only 10% of the relevance values and half of them on books they have not read. The vast majority of values come from single other forum members (3892 or 85%). Also, 3088 relevance values are based on judgements from members who have read the book (67%).

3.9 Evaluation

This year eight teams together submitted 34 runs (see Table 1). The official evaluation measure for this task is nDCG@10. It takes graded relevance values into account and concentrates on the top retrieved results. The results are

Table 8. Statistics on the types of member judgements on which the relevance values are based

	has read	not read	total
creator	218	220	438
other single	2666	1226	3892
other multi	204	38	242
total	3088	1484	4572

Table 9. Evaluation results for the official submissions. Best scores are in bold. Runs starting with * are manual runs

Group	Run	nDCG@10	P@10	MRR	MAP
RSLIS	run3.all-plus-query.all-doc-fields	**0.1361**	0.0653	0.2407	**0.1033**
UAms (ILLC)	inex13SBS.ti_qu.bayes_avg.LT_rating	0.1331	**0.0771**	**0.2437**	0.0953
UAms (ILLC)	inex13SBS.ti_qu_gr_na.bayes_avg...	0.1320	0.0668	0.2355	0.0997
RSLIS	run1.all-topic-fields.all-doc-fields	0.1295	0.0647	0.2290	0.0949
UAms (ILLC)	inex13SBS.ti_qu_gr_na	0.1184	0.0555	0.2169	0.0926
UAms (ILLC)	inex13SBS.ti_qu	0.1163	0.0647	0.2197	0.0816
ISMD	*run_ss_bsqstw_stop_words_free_member...	0.1150	0.0479	0.1930	0.0925
UAms (ILLC)	inex13SBS.qu.bayes_avg.LT_rating	0.1147	0.0661	0.2092	0.0794
ISMD	*run_ss_bsqstw_stop_words_free_2013	0.1147	0.0468	0.1936	0.0924
UAms (ILLC)	inex13SBS.ti.bayes_avg.LT_rating	0.1095	0.0634	0.2089	0.0772
ISMD	ism_run_ss_free_text_2013	0.1036	0.0426	0.1674	0.0836
ISMD	*run_ss_bsqstw_2013	0.1022	0.0416	0.1727	0.0830
ISMD	*run_ss_bsqstw_stop_words_free_query...	0.0940	0.0495	0.1741	0.0675
NTNU	aa_LMJM3	0.0832	0.0405	0.1537	0.0574
NTNU	az_LMJM3	0.0814	0.0376	0.1534	0.0504
NTNU	az_BM25	0.0789	0.0392	0.1517	0.0540
NTNU	aa_BM25	0.0780	0.0366	0.1426	0.0485
UAms (ILPS)	UAmsRetTbow	0.0664	0.0355	0.1235	0.0483
UAms (ILPS)	indri	0.0645	0.0347	0.1245	0.0445
NTNU	qa_LMD	0.0609	0.0345	0.1249	0.0396
LSIS/AMU	score_file_mean_R_2013_reranked	0.0596	0.0324	0.1200	0.0462
NTNU	qz_LMD	0.0577	0.0342	0.1186	0.0366
LSIS/AMU	score_file_SDM_HV_2013_reranked	0.0576	0.0292	0.1252	0.0459
LSIS/AMU	resul_SDM_2013	0.0571	0.0297	0.1167	0.0459
RSLIS	run2.query.all-doc-fields	0.0401	0.0208	0.0728	0.0314
CYUT	Run4.query.RW	0.0392	0.0287	0.0886	0.0279
CYUT	Run6.query.reviews.RW	0.0378	0.0284	0.0858	0.0244
CYUT	Run2.query.Rating	0.0376	0.0284	0.0877	0.0259
CYUT	Run1.query.content-base	0.0265	0.0147	0.0498	0.0220
CYUT	Run5.query.reviwes.content-base	0.0254	0.0153	0.0457	0.0209
CYUT	Run3.query.RA	0.0170	0.0087	0.0437	0.0166
OUC	sb_ttl_nar_10000_0.5	0.0100	0.0071	0.0215	0.0076
OUC	sb_ttl_nar_0.4	0.0044	0.0029	0.0104	0.0031
OUC	sb_ttl_nar_2500	0.0039	0.0032	0.0097	0.0032

shown in Table 9. None of the groups used user profile information for the runs they submitted. The best performing run is *run3.all-plus-query.all-doc-fields* by **RSLIS**, which used all topic fields combined against an index containing all available document fields. The second best group is **UAms (ILLC)** with run *inex13SBS.ti_qu.bayes_avg.LT_rating*, which uses only the topic titles and moderated query ran against an index containing the title information fields (title, author, edition, publisher, year), user-generated content fields (tags, reviews and awards) and the subject headings and Dewey decimal classification titles from the British Library and Library of Congress. The retrieval score of each book was then multiplied by a prior probability based on the Bayesian average of LT ratings for that book. The third group is **ISMD**, with manual run *run_ss_bsqstw_stop_words_free_member_free_2013* (to make the table fit on the page, it is shortened to *run_ss_bsqstw_stop_words_free_member...*). This run is generated after removing Book Search Query Stop Words (BSQSTW), standard stop words and the *member* field from the topics and running against an index where stop words are removed and the remaining terms are stemmed with the Krovetz stemmer. If we ignore the manual runs, ISMD is still the third group with the fully automatic run *ism_run_ss_free_text_2013*, which is generated using free text queries on Krovetz stemmed and stop words removed index.

Many teams used similar approaches, with query representations based on a combination of topic fields and indexes based on both professional and user-generated metadata. It seems that advanced models were implemented that combining topic statements with profile information or that treat professional metadata differently from user-generated content. This may be due to the late release the topics and profiles and the submission deadline being early because of changes in the schedule of CLEF. It is also possible that for most participants this task is felt to be a retrieval task modelled after standard TREC tasks, so there is little attention for recommendation aspects.

4 The Prove It (PI) Task

The goal of this task was to investigate the application of focused retrieval approaches to a collection of digitised 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 11 runs were submitted by 2 groups (6 runs by OUAC (ID=100) and 5 runs by University of Amsterdam (ID=4)), see Table 1.

4.1 The Digitized Book Corpus

The track builds on a collection of 50,239 out-of-copyright books[9], digitised 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.

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.

[9] Also available from the Internet Archive (although in a different XML format)

4.2 Topics

In recent years we have had a topic-base of 83 topics, and for 30 of them we have collected relevance judgments using crowdsourcing through Amazon Mechanical Turk [4, 6]. This year we added a new relevance criterium, namely appropriateness. It is not enough that a page confirms or refutes a fact, it should also come from a book that is trusted and of an appropriate genre. For a fact about Darwin's life, a famous biography on Darwin would be a more appropriate source than an obscure textbook on biology. Book pages are judged on two levels: 1) the extent to which a page confirms or refutes the factual claim, determined by how many of the atomic aspects of the claim are confirmed/refuted, and 2) the appropriateness of the book of the page is a part.

Aspect relevance Last year we introduced aspect relevance [6], where complex statements were broken down into atomic parts, which could be judge more easily individually. The overall relevance score of page for the whole statement would be the sum of relevance scores for the atomic parts.

To divide each topic into its primitive aspects (a process we refer to as "aspectisation") we developed a simple web-application with a database back-end, to allow anyone to aspectise topics. This resulted in 30 aspectised topics. The judgements were collected last year. For each page being assessed for confirmation / refutation of a topic, the assessor is presented with a user interface similar to Figure 2

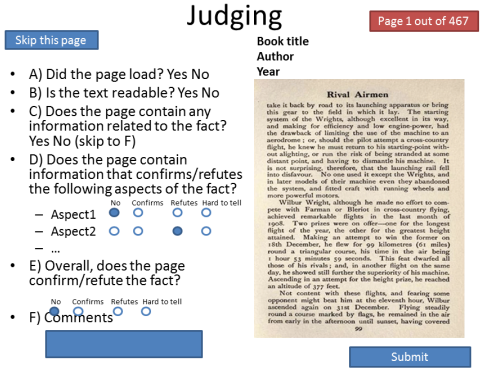

Fig. 2. An illustration of the planned assessment interface

This means that we go from a discrete (confirms / refute / none) assessment to a graded assessment, where a page may e.g. be assessed by a certain as 33 percent confirming a topic, if one of three aspects is judged as confirmed by him/her for that page. For the current assessment we have prepared 30 topics, for which the number of aspects range from 1 (very simple statements) to 6 per topic with an average of 2,83 aspects per topic.

Appropriateness For this year, the new assessment phase should establish whether the source books for the top-10 pages per topic are appropriate. At the time of writing, the assessment interface is not yet ready, but we expect to run this phase in the summer of 2013. The interface will present assessors with one of the top-10 pages, and provide them with the ability to browse through the rest of the book both via forward/backward buttons, an interactive table of contents in a sidebar as well as a page number box in which they can indicate to which specific page they want to jump. For each source book of the top-10 pooled pages, the assessor has to judge the appropriateness of the book. This appropriateness score is then propagated to all its pages that are part of the judgement pool.

4.3 Collected Relevance Assessments

As for the 2012 experiments, one hundred pages were pulled from the top-10 results of participant submissions for each topic. Assessments for each page and statement were collected from three assessors. New for this year, we also ask assessors to assess weather the book a presented page belongs to is appropriate for the task of confirming or refuting the statement. appropriateness is based on the book's genre or topic.

4.4 Evaluation Measures and Results

Result publication is awaiting the conclusion of the relevance assessment process.

5 The Structure Extraction (SE) Task

As in previous years, the goal of the SE task was to test and compare automatic techniques for extracting structure information from digitised books and building a hyperlinked table of contents (ToC). The task was motivated by the limitations of current digitisation and OCR technologies that produce the full text of digitised 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 2013, the task was run for the third time as a competition of the International Conference on Document Analysis and Recognition (ICDAR). Full details are presented in the corresponding specific competition description [3]. This year, the main novelty was that ground truthing was performed by an independent provider. This granted higher consistency and set participants free from ground truthing duties, which were a known drawback of participating to the task. The ground truth for the 2013 competition is already available online[10].

[10] https://doucet.users.greyc.fr/StructureExtraction/training/

Organization	Submitted runs	First submission
Elsevier	0	-
EPITA (France)	1	2013
INRIA (France)	0	-
Microsoft Development Center (Serbia)	1	2009
Nankai University (PRC)	1	2011
NII Tokyo (Japan)	0	-
University of Caen (France)	5	2009
University of Innsbruck (Austria)	1	2013
University of Würzburg (Germany)	1	2013

Table 10. Active participants of the Structure Extraction task.

Participation

Following the call for participation issued in January 2013, 9 organizations registered. As in previous competitions, several participants expressed interest but renounced due to time constraints. Of the 9 organizations that signed up, 6 submitted runs. This promising increase in active participants (6 out of 9), compared with previous years (4 out of 11), is likely a result of available training data and the removed obligation on creating ground truth ToCs.

Results

As in previous years [2], the 2013 task permitted to make manual annotations available to the community. The efforts of the 2013 round gave way to the addition of 967 new annotated book ToCs to the existing 1,037, nearly doubling the amount of available test data.

A summary of the performance of all the submitted runs is given in Table 11.

The Structure Extraction task was launched in 2008 to compare automatic techniques for extracting structure information from digitised books. While the construction of hyperlinked ToCs was originally thought to be a first step on the way to the structuring of digitised 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 Conclusions and plans

This was the third year of the Social Book Search Track. This year, the track ran three tasks: the Social Book Search task, the Prove It task and the Structure Extraction task.

RunID	Participant	F-measure
MDCS	MDCS	43.61%
Nankai	Nankai U.	35.41%
Innsbruck	Innsbruck U.	31.34%
Würzburg	Würzburg U.	19.61%
Epita	Epita	14.96%
GREYC-run-d	University of Caen	8.81%
GREYC-run-c	University of Caen	7.91%
GREYC-run-a	University of Caen	6.21%
GREYC-run-e	University of Caen	4.71%
GREYC-run-b	University of Caen	3.79%

Table 11. Summary of title-based performance scores for the Structure Extraction competition 2013 (F-measure for complete entries).

The Social Book Search (SBS) task continued which 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. The number of active participants has doubled from 4 to 8, suggesting a promising future for the task.

Expanding on the evaluation of the previous two years, we delved deeper into the nature of book search information needs and book suggestions from the LT forums. We annotated both 1) the information needs described by the starters of topic thread there were asking for book suggestions, and 2) the books suggested by LT members in the thread.

We found that most social book search topics have requirements related to the content of the book, such as topic and genre, but that metadata, familiarity and engagement—asking for books by a certain author, books that are similar to a particular (set of) book(s) and books that fit a certain mood, interest or quality respectively—are also important aspects. This strengthens and extends our findings from last year that social book search topics express complex needs that are hard to satisfy with current book search services, but also to specific for typical recommendation systems.

Another finding in the SBS task is that forum members mostly suggest books they have read although there are also many cases where it is hard to judge from what they write about the books they suggest. When it is clear they have read the books they suggest read, they are mostly positive, supporting our choice of using forum suggestions as relevance judgements. When they suggest books they have not read, or when it is hard to tell, their are often neutral. This could be a signal that suggestions of unread books are closer to traditional topical relevance judgements and suggestions of read books are topic specific recommendations that satisfy all or most of the complex combination of relevance aspects.

The evaluation has shown that the most effective systems incorporate the full topic statement, which includes the title of the topic thread, a query provided by the annotator, the name of the forum discussion group, and the full first

message that elaborates on the request. However, the best system is a plain full-text retrieval system that ignores all user profile information. It could be that the suggestions by members other than the topic starter favour non-personalised retrieval models and thereby muddle the personalised signal of suggestions supported by the topic starter.

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.

This year the Prove It task changed somewhat by adding the criterium that returned pages should come from reliable sources. That is, authoritative books of the appropriate genre. The assessment phase is yet to start so there are no evaluation results yet. Due to the low number of participants over the last years, the task will probably not run again next year.

The SE task relies on a subset of the 50,000 digitised books of the PI task. In 2013, the participants were to extract the tables of contents of 1,000 books extracted from the whole PI book collection. In previous years, the ground truth was constructed collaboratively by participating institution. For the first, time in 2013 the ground truth production was performed by an external provider. This centralised construction granted better consistency. In addition, it also validated the collaborative process used since 2009, as the results this year were in line with those of the previous rounds.

The structure extraction task has reached a record high number of active participants, and has for the first time witnessed an improvement of the state of the art. In future years, we aim to investigate the usability of the extracted ToCs, both for readers in navigating books and systems that index and search parts of books. To be able to build even larger evaluation sets, we hope to experiment with crowdsourcing methods. This may offer a natural solution to the evaluation challenge posed by the massive data sets handled in digitised libraries.

Acknowledgments We are very grateful to Toine Bogers for helping us with the topic annotation tool and for recruiting LIS students to be annotators for this year's topic selection and relevance assessments.

Bibliography

[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] A. Doucet, G. Kazai, B. Dresevic, A. Uzelac, B.Radakovic, and N. 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.

[3] A. Doucet, G. Kazai, S. Colutto, and G. Mühlberger. Overview of the ICDAR 2013 Competition on Book Structure Extraction. In *Proceedings of the Twelfth International Conference on Document Analysis and Recognition (ICDAR'2013)*, page 6, Washington DC, USA, August 2013.

[4] G. Kazai, J. Kamps, M. Koolen, and N. 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.

[5] M. Koolen, J. Kamps, and G. Kazai. Social Book Search: The Impact of Professional and User-Generated Content on Book Suggestions. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM 2012)*. ACM, 2012.

[6] M. Koolen, G. Kazai, J. Kamps, M. Preminger, A. Doucet, and M. Landoni. Overview of the INEX 2012 social book search track. In S. Geva, J. Kamps, and R. Schenkel, editors, *Focused Access to Content, Structure and Context: 11th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'12)*, LNCS. Springer, 2013.

[7] K. Reuter. Assessing aesthetic relevance: Children's book selection in a digital library. *JASIST*, 58(12):1745–1763, 2007.