Overview of the INEX 2014 Social Book Search Track

Marijn Koolen[1], Toine Bogers[2], Gabriella Kazai[2], Jaap Kamps[1], and Michael Preminger[3]

[1] University of Amsterdam, Netherlands
{marijn.koolen,kamps}@uva.nl
[2] Aalborg University Copenhagen
toine@hum.aau.dk
[3] Microsoft Research, United Kingdom
a-gabkaz@microsoft.com
[4] Oslo and Akershus University College of Applied Sciences, Norway
michaelp@hioa.no

Abstract. The goal of the INEX 2014 Social Book Search Track is to evaluate approaches for supporting users in searching collections of books based on book metadata and associated user-generated content. The track investigates the complex nature of relevance in book search and the role of traditional and user-generated book metadata in retrieval. We extended last year's investigation into the nature of book suggestions from the LibraryThing forums and how they compare to book relevance judgements. Participants were encouraged to incorporate rich user profiles of both topic creators and other LibraryThing users to explore the relative value of recommendation and retrieval paradigms for book search. 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.

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 two 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,

Table 1. Active participants of the INEX 2014 Social Book Search Track and number of contributed runs

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

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

The 2014 edition represents the fourth consecutive year the SBS task has run and oncemore the test collection used is the Amazon/LibraryThing collection of 2.8 million documents. LibraryThing forum requests for book suggestions, combined with annotation of these requests resulted in a topic set of 680 topics with graded relevance judgments. Compared to 2013, there are three important changes: (1) a much larger set of 94,000+ user profiles was provided to the participants this year; (2) an additional 300 forum topics were annotated, bringing the total number of topics up to 680; and (3) the *Prove It* task did not run this year.

In this paper, we report on the setup and the results of the SBS Track at the 2014 INEX@CLEF Lab. First, in Section 2, we give a brief summary of the participating organisations. The SBS task itself is described in Section 3. Sections 4 and 5 describe the test collection and the evaluation process in more detail. We close in Section 6 with a summary and plans for 2014.

2 Participating Organisations

A total of 64 organisations registered for the track (compared with 68 in 2013, 55 in 2012 and 47 in 2011). At the time of writing, we counted 8 active groups (compared with 8 in 2013, 5 in 2012 and 10 in 2011), see Table 1.

3 Social Book Search Task Setup

3.1 Track Goals and Background

The goal of the Social Book Search (SBS) track is to evaluate the value of professional metadata and user-generated content for book search on the Web and 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 [3], 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 [2] or recommendation.

The SBS track 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. 0 The track 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 track 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?

3.2 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 of obtaining 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.

3.3 Task description

The task is to reply to a user request posted on a LT forum (see Section 4.1) 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 track 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 track 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.

3.4 Submission Format

Participants are asked to return a ranked list of books for each user query, ranked by order of relevance, where the query is described in the LT forum thread. We adopt the submission format of TREC, with a separate line for each retrieval result (i.e., book), 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 4.1). For the other five runs, participants could use any field in the topic statement.

4 Test Collection

We use and extend the Amazon/LibraryThing (A/LT) corpus crawled by the University of Duisburg-Essen for the INEX Interactive Track [1]. The corpus contains a large collection of book records with controlled subject headings and classification codes as well as social descriptions, such as tags and reviews. 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 the

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	

subject descriptors are noisy, with a number of inappropriately assigned descriptors that seem unrelated to the books.

Each book is identified by an ISBN. Note that 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%).

[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/

Although there is no single library catalogue that covers all books available on 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.

4.1 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, in turn, 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 valuable judgements on the relevance of books. Additional relevance information can be gleaned from the discussions on the threads. Consider, for example, topic 129939[7]. 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, 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.

[7] URL: http://www.librarything.com/topic/129939

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

In the following, we first describe the topic selection and annotation procedure, then how we used the annotations to assign relevance values to the suggestions, and finally the user profiles, which were then provided with each topic.

Topic selection Over the past two years, we had a group of eight different Information Science students annotate the narratives of a random sample of 2,646 LT forum topics. Three of these students hailed from the Royal School of Library and Information Science in Copenhagen, three from the Oslo & Akershus University of Applied Sciences, and one from Aalborg University Copenhagen.

We created a Web interface to help our annotators (1) identify topic threads as either *book requests* (describing a valid information need) or *non-requests* (covering any other type of discussion topic); (2) annotate the selected book search 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 other LT members in the thread. This latter task included questions on whether the suggesters appear to have read the suggested books and what their attitudes seem to be towards the books, i.e., whether their recommendation is positive, negative or neutral.

Of the 2,646 topics annotated by the students, 944 topics (36%) were identified as containing a book search information need. Because we want to investigate the value of recommendations, we use only topics where the topic creators add books to their catalogue both before (pre-catalogued) and after starting the topic (post-catalogued). Without the former, recommender systems have no profile to

Table 3. Distribution of relevance aspects over the annotated requests. The left side of the table displays the distribution of relevance aspects over the 680 topics. The right side of the table shows the distribution of the number of aspects expressed in a single topic.

Aspect	**#**	**%**	**# aspects**	**# topics**	**%**
Accessibility	106	16	1	191	28
Content	523	77	2	260	38
Engagement	154	23	3	183	27
Familiarity	261	38	4	37	5
Known-item	97	14	5	7	1
Metadata	177	26	6	2	0
Novelty	29	4			
Socio-Cultural	108	16			
Total	680	100		680	100

work with and without the latter the recommendation part cannot be evaluated. This leaves 680 topics. These topics were combined with all the pre-catalogued books of the topic creators' profiles and distributed to participating groups.

Topics can represent complex information needs, often 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 [3] 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 books 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 annotation 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 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 3. Book search information needs on the LT forums almost always (77% of the 680 topics) contain content aspects. This reinforces the traditional choice of designing best-match retrieval models around aspects of document content. Metadata aspects, such as book title and author, are present in 26% of the data set. Other important aspects are *familiarity* (38%) and *engagement* (23%). 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.

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. The latter was obtained in order to provide queries that 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 680 topics, 306 (45%) asked for suggestions on fiction books, 122 (18%) on non-fiction, 95 (14%) on both fiction and non-fiction, and for 157 topics (23%) the annotator could not tell.

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>
```

```
<catalog>
  <book>
    <LT_id>9036</LT_id>
    <entry_date>2007-09</entry_date>
    <rating>0.0</rating>
    <tags></tags>
  </book>
  <book>
    ...
```

Finally, annotators had to label each touchstone provided by LT members (including any provided by the topic starter). They had to indicate whether the suggester *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 suggester 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.

In the majority of cases (61%) members suggested books that they have read. It is rather rare for suggesters to state that they have not read a suggested book (8%). More often, suggesters do not reveal whether they have read the book or not (28%). Books mentioned in response to a book search request are often presented in a positive (47%) or neutral (39%) way. Both positive and negative suggestions tend to come from members who have read the books (71% and 87% respectively). When books are mentioned in a neutral way, it is often difficult to tell whether the book has been read by the suggester, although a third of the neutral mentions comes from members who have read the book.

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.

Operationalisation of forum judgement labels The annotated suggestions 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.

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 a 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 there is no clear difference in the value of such recommendations.
- When the annotator was *not sure* if a suggestion was 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 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. catalogued by topic creator
   - 1.1 post-catalogued → $rv = 8$
   - 1.2 pre-catalogued → $rv = 0$
2. single judgement
   - 2.1 starter has read judgement → $rv = 0$

2.2 starter has not read judgement
2.2.2 starter positive → $rv = 8$
2.2.2 starter neutral → $rv = 2$
2.2.2 starter negative → $rv = 0$
2.3 other member has read judgement
2.3.3 has read positive → $rv = 4$
2.3.3 has read neutral → $rv = 2$
2.3.3 has read negative → $rv = 0$
2.4 other member has not read judgement
2.4.4 not read positive → $rv = 3$
2.4.4 not read neutral → $rv = 2$
2.4.4 not read negative → $rv = 0$

3. multiple judgements
3.1 multiple has read judgements
3.1.1 some positive, no negative → $rv = 6$
3.1.1 #positive > #negative → $rv = 4$
3.1.1 #positive == #negative → $rv = 2$
3.1.1 all neutral → rv=2
3.1.1 #positive < #negative → $rv = 1$
3.1.1 no positive, some negative → $rv = 0$
3.2 multiple not read judgements
3.2.2 some positive, no negative → $rv = 4$
3.2.2 #positive > #negative → $rv = 3$
3.2.2 #positive == #negative → $rv = 2$
3.2.2 all neutral → rv=2
3.2.2 #positive < #negative → $rv = 1$
3.2.2 no positive, some negative → $rv = 0$

This results in graded relevance values with seven possible values (0, 1, 2, 3, 4, 6, 8).

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 creators and other LT users, which contain information on which books they have in their personal catalogue on LT, which ratings and tags they assigned to them and a social network of friendship relations, interesting library relations and group memberships. These profiles may provide important signals on the user's topical and genre interests, reading level, which books they already know and which ones they like and don't like. These profiles were scraped from the LT site, anonymised and made available to participants. This allows Track participants to experiment with combinations of retrieval and recommender systems. 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 when selecting books. This may partly be because she wants to explore different

Table 4. User profile statistics of the topic creators and all other users.

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

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.

To anonymise all user profiles, we first removed all friendship and group membership connections and replaced the user name with a randomly generated string. The cataloguing date of each book was reduced to the year and month. What is left is an anonymised user name, book ID, month of cataloguing, rating and tags.

Basic statistics on the number of books per user profile is given in Table 4. By the time users ask for book recommendations, most of them already have a substantial catalogue (pre-catalogued). The distribution is skewed, as the mean (587) is higher than the median (239). After posting their topics, users tend to add many more books (post-catalogued), but fewer than they have already added. Compared to the other users in our crawl (median of 134 books), the topic creators are the more active users, with larger catalogues (median of 432 books).

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.[8]

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.

5 Evaluation

This year, eight teams submitted a total of 40 runs (see Table 1). The official evaluation measure for this task is nDCG@10. It takes graded relevance values into account and is designed for evaluation based on the top retrieved results. In addition, P@10, MAP and MRR scores will also be reported, with the evaluation results shown in Table 5.

None of the best-performing groups used user profile information for the runs they submitted. The best performing run is *run6.SimQuery1000.rerank_all.L2R_RandomForest* by USTB, which used all topic fields combined against an index containing all available document fields. The run is re-ranked with 12 different re-ranking strategies, which are then combined adaptively using learning-to-rank. The second group is UJM with run *326*, which uses BM25 on the title, mediated query and narrative fields, with the parameters optimised for the narrative field. The third group is **lsis**, with *InL2*. This run is based on the InL2 model, the index is built from all fields in the book xml files. The system uses the mediated query, group and narrative fields as a query.

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

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

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

Table 5. Evaluation results for the official submissions. Best scores are in bold. Runs marked with * are manual runs.

Group	Run	nDCG@10	P@10	MRR	MAP	Profiles
USTB	run6.SimQuery1000.rerank_all.L2R_RandomForest	**0.303**	**0.464**	**0.232**	0.390	No
USTB	run4.newXml.rerank_all.L2R_RandomForest	0.142	0.258	0.102	0.390	No
UJM	326	0.142	0.275	0.107	0.426	No
USTB	run3.newXml.rerank_all.L2R_Coordinate	0.138	0.256	0.101	0.390	No
USTB	run5.newXml.rerank_all.L2R_RankNet	0.133	0.246	0.098	0.390	No
USTB	run2.newXml.rerank_T	0.131	0.246	0.096	0.390	No
USTB	run1.newXml.feedback	0.128	0.246	0.095	0.390	No
LSIS	InL2	0.128	0.236	0.101	0.441	No
AAU	run1.all-plus-query.all-doc-fields	0.127	0.239	0.097	**0.444**	No
AAU	run3.all-plus-query.all-doc-fields	0.120	0.227	0.090	0.425	No
CYUT	Type2QTGN	0.119	0.246	0.086	0.340	No
CYUT	0.95AverageType2QTGN	0.119	0.243	0.085	0.332	No
UJM	328	0.117	0.226	0.088	0.392	Yes
UJM	329	0.116	0.217	0.087	0.392	Yes
UJM	325	0.115	0.214	0.087	0.392	Yes
LSIS	InL2Feedback	0.114	0.230	0.094	0.434	No
UJM	324	0.112	0.214	0.086	0.392	No
LSIS	InL2tagFeedback	0.102	0.212	0.075	0.388	No
UvA	inex14.ti_qu.fb.10.50.5000	0.097	0.179	0.073	0.421	No
UMD	Full_TQG_fb.10.50_0.0000227_50	0.097	0.188	0.069	0.328	Yes
UMD	Social_TQG_fb.10.50_0.0000222_50	0.096	0.184	0.067	0.327	Yes
UMD	Full_TQG_fb.10.50_0.0000255_100	0.096	0.188	0.068	0.328	Yes
UvA	inex14.ti_qu_gr.fb.10.50.5000	0.095	0.162	0.074	0.436	No
UvA	inex14.ti_qu.5000	0.095	0.173	0.073	0.412	No
UMD	Full_TQG_fb.10.50_traditional	0.095	0.185	0.068	0.328	No
UvA	inex14.ti_qu_gr.5000	0.094	0.163	0.074	0.418	No
UMD	Full_TQ_fb.10.50_0.0000247_100	0.092	0.176	0.064	0.321	Yes
UMD	Full_T_fb.10.50_0.0000260_100	0.070	0.139	0.047	0.253	Yes
*ISMD	354	0.067	0.123	0.049	0.285	No
LSIS	sdm_Rating	0.062	0.120	0.047	0.314	No
LSIS	sdm_concept	0.056	0.118	0.039	0.253	No
*ISMD	341	0.056	0.106	0.042	0.236	No
LSIS	sdm_tag_feedback	0.055	0.112	0.040	0.267	No
UJM	345	0.052	0.113	0.037	0.383	Yes
*ISMD	350	0.048	0.090	0.036	0.211	No
AAU	run2.query.all-doc-fields	0.047	0.090	0.035	0.304	No
*ISMD	355	0.038	0.089	0.026	0.124	No
CYUT	0.95RatingType2QTGN	0.034	0.101	0.021	0.200	No
CYUT	0.95WRType2QTGN	0.028	0.084	0.018	0.213	No
*ISMD	342	0.010	0.018	0.007	0.081	No

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

6 Conclusions and Plans

This was the fourth year of the Social Book Search Track. The track ran only a single tasks: the system-oriented Social Book Search task, which continued its focus on both the relative value of professional and user-generated metadata and the retrieval and recommendation aspects of the LT forum users and their information needs. The number of active participants remained stable at 8, suggesting there is still significant interest the task.

Expanding on the evaluation of the previous year, we kept the evaluation procedure the same, but included larger sets of topics and user profiles.

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. Social book search topics express complex needs that are hard to satisfy with current book search services, but that are also too specific for typical recommendation systems.

Forum members mostly suggest books they have read, although there are also many cases where it is hard to judge based on what they write about their suggestions. When it *is* clear that they themselves have read their suggestions, they are mostly positive, which lends supports for our choice to using forum suggestions as relevance judgements. When they suggest books they have not read themselves—or when it is hard to tell—they are often neutral in their descriptions. 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 and the full first message that elaborates on the request. For many of the top performing systems the parameters have been optimised through training. The best-performing systems uses learning-to-rank to combine multiple re-ranking methods.

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.

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] 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.

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