Overview of the SBS 2015 Suggestion Track

Marijn Koolen[1], Toine Bogers[2], and Jaap Kamps[1]

[1] University of Amsterdam, Netherlands
{marijn.koolen,kamps}@uva.nl
[2] Aalborg University Copenhagen
toine@hum.aau.dk

Abstract. The goal of the SBS 2015 Suggestion Track is to evaluate approaches for supporting users in searching collections of books who express their information needs both in a query and through example books. 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. In terms of systems evaluation, the most effective systems include some form of learning-to-rank. It seems that the complex nature of the requests and the book descriptions, with multiple sources of evidence, requires a careful balancing of system parameters.

1 Introduction

The goal of the Social Book Search 2015 Suggestion Track[3] is to investigate techniques to support users in searching for books in catalogues of professional metadata and complementary social media. 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,

The *Social Book Search* (SBS) 2015 Suggestion Track, framed within the scenario of a user searching a large online book catalogue for a given topic of

[3] See http://social-book-search.humanities.uva.nl/#/suggestion

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 2015 Suggestion Track is a continuation of the INEX SBS Track that ran from 2011 up to 2014. For this fifth edition the focus is on search requests that combine a natural language description of the information need as well as example books, combining traditional ad hoc retrieval with query-by-document. The information needs are derived from the LibraryThing (LT) discussion forums. LibraryThing forum requests for book suggestions, combined with annotation of these requests resulted in a topic set of 208 topics with graded relevance judgments. A test collection is constructed around these information needs and the Amazon/LibraryThing collection, consisting of 2.8 million documents. The Suggestion Track runs in close collaboration with the SBS Interactive Track,[4] which is a user-centered track where interfaces are developed and evaluated and user interaction is analysed to investigate how book searchers make use of professional metadata and user-generated content.

In this paper, we report on the setup and the results of the 2015 Suggestions Track as part of the SBS Lab at CLEF 2015. 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 2016.

2 Participating Organisations

A total of 25 organisations registered for the track (compared with 64 in 2014, 68 in 2013, 55 in 2012 and 47 in 2011). Although the number of registered teams has dropped, the number of active teams has increased from 8 in 2014 to 11 in 2015, 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

[4] See http://social-book-search.humanities.uva.nl/#/interactive

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

Institute	Acronym	Runs
Aalborg University Copenhagen	AAU	1
Aix-Marseille Université CNRS	LSIS	6
Chaoyang University of Technology	CSIE	4
Laboratoire d'Informatique de Grenoble	MRIM	6
Laboratoire Hubert Curien, Université de Saint-Etienne	LaHC	6
Oslo & Akershus University College of Applied Sciences	Oslo_SBS	4
Research Center on Scientific and Technical Information	CERIST	4
University of Amsterdam	UvA	3
Université de Neuchâtel, Institut de Recherche en Informatique de Toulouse	MIIB	6
University of Jordan	IR@JU	2
University of Science and Technology Beijing	USTB_PRIR	6
Total		48

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 (Reuter, 2007), 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 of 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 (Koolen et al., 2012) 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.

The track builds on the INEX Amazon/LibraryThing (A/LT) collection (Beckers et al., 2010), 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 Suggestion 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 Suggestion 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 2015 Suggestion Task

This year, the task focuses on search requests that combine a rich narrative description of the information need and one or more example books that the requester considers positive or negative. The challenge for systems is to find the right balance between the two types of evidence and how to use the natural language statement to infer the relevant aspects of the example books.

3.5 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 (Beckers et al., 2010). 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 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

[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	

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 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

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

a large collection of online book records. We use a sample of these forum topics to evaluate systems participating in the Suggestion Track.

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.

In the following, we first describe the topic selection and annotation procedure, then how we used the annotations to assign relevance values to the

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

Table 3. Number of examples per topic

	N	Total	Min	Max	Median	Mean	St.dev
Examples per topic	208	516	1	21	2	2.48	2.38

suggestions, and finally the user profiles, which were then provided with each topic.

Topic selection The topic set of 2015 is a subset of the 2014 topic set, focusing on topics with both a narrative description of the information need and one or more example books to guide the suggestions.

In 2013 and 2014, we had a group of eight different Information Science students annotate the narratives of a random sample of 2,646 LT forum topics (Koolen et al., 2013, 2014). Of the 2,646 topics annotated by the students, 944 topics (36%) were identified as containing a book search information need. Because we want to investigate the value of recommendations, we use only topics where the topic creators add books to their catalogue both before (pre-catalogued) and after starting the topic (post-catalogued). Without the former, recommender systems have no profile to work with and without the latter the recommendation part cannot be evaluated. Finally, we select only those topics where the request contains explicit mentions (marked up in *touchstones*) of books that function as examples of what the requester is looking for, or that have some aspects that the requester does not want. This leaves 208 topics for the 2015 topic set. These topics were combined with all the pre-catalogued books of the topic creators' profiles and distributed to participating groups.

Each topic has at least one example book provided by the requester that helps other forum members understand in which direction the requester is thinking. The number of examples ranges from 1 to 21 (Table 3), with a median and mean of 2 and 2.48 respectively. Further, annotators indicated whether an example book was given as a positive example—i.e. they are looking for something along the lines of the example—or as a negative example, where the example is broadly relevant but has aspects that the requester does not want in the suggested books.

After annotation, the topic in Figure 1 (topic 99309) is distributed to participants in the following format:

```
<topic id="99309">
  <query>Politics of Multiculturalism</query>
  <title>Politics of Multiculturalism Recommendations?</title>
  <group>Political Philosophy</group>
  <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>
<examples>
  <example>
    <LT_id>164382</LT_id>
    <hasRead>yes</hasRead>
    <sentiment>neutral</sentiment>
  </example>
</examples>
<catalog>
  <book>
    <LT_id>9036</LT_id>
    <entry_date>2007-09</entry_date>
    <rating>0.0</rating>
    <tags></tags>
  </book>
  <book>
    ...
```

The hyperlink markup, represented by the `<a>` tags, is added by the *Touchstone* technology of LT. The rest of the markup is generated specifically for the Suggestion Track. Above, the example book with *LT_id* 164382 is annotated as an example that the requester is neutral about. It has positive and negative aspects. From the request, forum members can understand how to interpret this example.

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.

In addition to the explicitly marked up books, e.g., the examples and suggestions, we noticed that there are other book titles that are not marked up but are intended as suggestions. In some cases this is because the suggester is not aware of the *Touchstone* syntax or because it fails to identify the correct book and they cannot manually correct it. To investigate the extent of this issue and to make the list of identified suggestions more complete, in 2015 we manually labeled all suggested books that were not marked up by *Touchstone* in each forum thread of the 208 topics.

This resulted in 830 new suggestions (a mean of 4 per topic). From the touchstones we extracted 4240 suggestions (20.4 per topic), so the manually extracted suggestions bring the total to 5070 (24.4), an increase of 20%. Multiple user may suggest the same books, so the total number of suggested books is lower. The 4240 touchstone suggestion represent 3255 books (15.6 per topic). With the manually extracted suggestions, this increases to 3687 (17.7 per topic), an increase of 13%. The newly added suggestions therefore increase the recall base but also increase the number of recommendations for some of the touchstone suggestions.

Operationalisation of forum judgement labels The mapping from annotated suggestions to relevance judgements uses the same process as in 2014. Note that some of the books mentioned in the forums are not part of the 2.8 million books in our collection. These suggestions removed from the suggestions any books that are not in the INEX A/LT collection. The numbers reported in the previous section were calculated after this filtering step.

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.
- A work with only negative suggestions has no value for the requester when found in the search results.
- *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 a single forum member mentions 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. To determine what relevance values to use, we observe that there are positive, neutral and negative suggestions by one or multiple suggesters. Based on the simplifying assumption that a work that is only mentioned negatively has no value for the suggester when found in the search results ($rv = 0$), we expect that works with more negative than positive suggestions have at least some value, but less than works with on average either neutral suggestions or positive suggestions. Therefore, a work with on average negative suggestions has the lowest positive relevance value $rv = 1$. On average neutral suggestions are the next level, with $rv = 2$. Works with on average positive suggestions get a relevance value higher than two, with a single positive suggestion or a mix of positive and negative suggestion getting an additional relevance point ($rv = 3$) and multiple positive suggestions two additional points ($rv = 4$). If the judges have read the books, the additional relevance points are

multiplied by two because they represent more reliable judgements. Specifically, the values are assigned according to the following scheme:

1. catalogued by topic creator
   1.1 post-catalogued $\rightarrow rv = 8$
   1.2 pre-catalogued $\rightarrow rv = 0$
2. single judgement
   2.1 starter has read judgement $\rightarrow rv = 0$
   2.2 starter has not read judgement
      2.2.2 starter positive $\rightarrow rv = 8$
      2.2.2 starter neutral $\rightarrow rv = 2$
      2.2.2 starter negative $\rightarrow rv = 0$
   2.3 other member has read judgement
      2.3.3 has read positive $\rightarrow rv = 4$
      2.3.3 has read neutral $\rightarrow rv = 2$
      2.3.3 has read negative $\rightarrow rv = 0$
   2.4 other member has not read judgement
      2.4.4 not read positive $\rightarrow rv = 3$
      2.4.4 not read neutral $\rightarrow rv = 2$
      2.4.4 not read negative $\rightarrow rv = 0$
3. multiple judgements
   3.1 multiple has read judgements
      3.1.1 some positive, no negative $\rightarrow rv = 6$
      3.1.1 #positive > #negative $\rightarrow rv = 4$
      3.1.1 #positive == #negative $\rightarrow rv = 2$
      3.1.1 all neutral $\rightarrow$ rv=2
      3.1.1 #positive < #negative $\rightarrow rv = 1$
      3.1.1 no positive, some negative $\rightarrow rv = 0$
   3.2 multiple not read judgements
      3.2.2 some positive, no negative $\rightarrow rv = 4$
      3.2.2 #positive > #negative $\rightarrow rv = 3$
      3.2.2 #positive == #negative $\rightarrow rv = 2$
      3.2.2 all neutral $\rightarrow$ rv=2
      3.2.2 #positive < #negative $\rightarrow rv = 1$
      3.2.2 no positive, some negative $\rightarrow rv = 0$

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

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	208	135,722	1	5884	270	653	991
Post-catalogued	208	74,240	1	5619	146	357	587
Total catalogue	208	209,962	2	6272	541	1009	1275
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

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 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 (653) is higher than the median (270). 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 135 books), the topic creators are the more active users, with larger catalogues (median of 541 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, 11 teams submitted a total of 48 automatic runs (see Table 1) and one manual run. We omit the manual run, as it is a ranking of last year's Qrels. The official evaluation measure for this task is nDCG@10. It takes graded relevance values into account and is designed for evaluation based on the top retrieved results. In addition, P@10, MAP and MRR scores will also be reported, with the evaluation results shown in Table 5.

The best runs of the top 5 groups are described below:

1. *MIIB - Run6* (rank 1): For this run, queries are generated from all topic fields and applied on a BM25 index with all textual document fields merged into a single field. A Learning-to-rank framework is applied using random forest on 6 result lists as well as the price, the book length and the ratings. Results are re-ranked based on tags and ratings.
2. *CERIST - CERIST_TOPICS_EXP_NO* (rank 2): The terms of topics have been combined with the top tags extracted from the example books mentioned in the book search request then the BM15 model has been used to rank books. The books which have been catalogued by the users have been removed.

[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.

Rank	Group	Run	nDCG@10	P@10	MRR	MAP	Profiles
1	MIIB	Run6	0.186	0.394	0.105	0.374	no
2	CERIST	CERIST_TOPICS_EXP_NO	0.137	0.285	0.093	0.562	yes
3	MIIB	Run2	0.130	0.290	0.074	0.374	no
4	CERIST	CERIST_TOPICS_EXP	0.113	0.228	0.080	0.558	yes
5	USTB_PRIR	run5-Rerank-RF-example	0.106	0.232	0.068	0.365	no
6	MRIM	LIG_3	0.098	0.189	0.069	0.514	yes
7	MRIM	LIG_2	0.096	0.185	0.069	0.514	no
8	MIIB	Run5	0.095	0.235	0.062	0.374	no
9	MRIM	LIG_4	0.095	0.181	0.068	0.514	yes
10	MIIB	Run4	0.094	0.232	0.061	0.375	no
11	MIIB	Run3	0.094	0.237	0.062	0.374	no
12	CERIST	CERIST_TOPICS	0.093	0.204	0.066	0.497	yes
13	MRIM	LIG_5	0.093	0.179	0.067	0.515	yes
14	MRIM	LIG_6	0.092	0.174	0.067	0.513	yes
15	CERIST	CERIST_EXAMPLES	0.090	0.189	0.060	0.448	yes
16	MRIM	LIG_1	0.090	0.173	0.063	0.508	no
17	LaHC_Saint-Etienne	UJM_2	0.088	0.174	0.065	0.483	no
18	USTB_PRIR	run4-Rerank-RF	0.088	0.189	0.056	0.359	no
19	AAU	allfields-jm	0.087	0.191	0.061	0.420	yes
20	LaHC_Saint-Etienne	UJM_6	0.084	0.160	0.060	0.483	no
21	MIIB	Run1	0.082	0.189	0.054	0.375	no
22	Oslo_SBS_iTrack_group	baseLine	0.082	0.182	0.052	0.341	no
23	CSIE	0.95AverageType2QTGN	0.082	0.194	0.050	0.319	no
24	LaHC_Saint-Etienne	UJM_1	0.081	0.167	0.056	0.471	no
25	LSIS-OpenEdition	INL2_SDM_Graph_LSIS	0.081	0.183	0.058	0.401	no
26	CSIE	Type2QTGN	0.080	0.191	0.052	0.325	no
27	Oslo_SBS_iTrack_group	sortedPace	0.080	0.182	0.051	0.341	no
28	USTB_PRIR	run3-UpperNar-abs-ex	0.079	0.197	0.052	0.312	no
29	LaHC_Saint-Etienne	UJM_3	0.079	0.155	0.059	0.485	no
30	LaHC_Saint-Etienne	UJM_4	0.079	0.158	0.055	0.471	no
31	LSIS-OpenEdition	INL2_fdep_SDM_LSIS	0.076	0.171	0.057	0.401	no
32	LSIS-OpenEdition	INL2_fdep_Graph_LSIS	0.075	0.162	0.054	0.388	no
33	LaHC_Saint-Etienne	UJM_5	0.074	0.150	0.054	0.471	no
34	LSIS-OpenEdition	INL2_fulldep_LSIS_OE	0.070	0.155	0.052	0.388	no
35	LSIS-OpenEdition	INL2_Gph_SimJac_LSIS	0.069	0.158	0.052	0.393	no
36	LSIS-OpenEdition	INL2_SelectDep_LSIS	0.069	0.161	0.053	0.382	no
37	UAmsterdam	UAmsQTG_KNN_L.070	0.068	0.160	0.051	0.388	yes
38	UAmsterdam	UAmsQTG_L1.00	0.065	0.140	0.050	0.341	no
39	USTB_PRIR	run2-Upper_narrative-abstract	0.061	0.155	0.042	0.309	no
40	USTB_PRIR	run1-example	0.042	0.120	0.022	0.029	no
41	CSIE	0.95RatingType2QTGN	0.032	0.113	0.019	0.214	no
42	CSIE	0.95WRType2QTGN	0.023	0.072	0.015	0.216	no
43	Oslo_SBS_iTrack_group	pace_1.2	0.012	0.042	0.009	0.254	no
44	Oslo_SBS_iTrack_group	pace_1.3	0.012	0.043	0.009	0.247	no
45	IR@JU	KASIT_1	0.011	0.023	0.006	0.009	no
46	IR@JU	KASIT_2	0.010	0.021	0.004	0.010	no
47	UAmsterdam	UAmsKNN_L0.00	0.006	0.020	0.004	0.139	yes

3. *USTB_PRIR - run5-Rerank-RF-example* (rank 5): This run is a mixture of two runs (*run1-example* and *run4-Rerank-RF*). The former ranks the example books for each topic. The latter is a complex run based on re-ranking with 11 strategies and learning-to-rank with random forest.
4. *MRIM - LIG_3* (rank 6): This run is a weighted linear fusion of a BM25F run on all fields, an LGD run on all fields, and the topic profile (from top tf terms of books in catalog), and the two "best friends" profiles according to similarity of marks on books.
5. *LaHC_Saint-Etienne - UJM_2* (rank 17): This run is based on the Log Logistic LGD model, with an index based on all document fields. For retrieval, the query is constructed from the title, mediated query, group and narrative fields in the topic statement.

Most of the top performing systems, including the best (MIIB's *Run6*) make no use of user profile information. There are 11 systems that made use of the user profiles, with 4 in the top 10 (at ranks 2, 4, 6 and 9). So far, the additional value of user profiles has not been established. The best systems combine various topic fields, with parameters trained for optimal performance. Several of the best performing systems make use of learning-to-rank approaches, suggesting book search is a domain where systems need to learn from user behaviour what the right balance is for the multiple and diverse sources of information, both from the collection and the user side.

6 Conclusions and Plans

This was the first year of the SBS Suggestion Track, which is a continuation from the SBS Track at INEX 2011–2014. The overall goal remains to investigate the relative value of professional metadata, user-generated content and user profiles, but the specific focus for this year is to construct a test collection to evaluate systems dealing with complex book search requests that combine an information need expressed in a natural language statement and through example books. The number of active participants increased to 11, suggesting this specific focus of interest in the IR community.

Extended the setup of the previous year, we kept the evaluation procedure the same, but included manually extracted suggestions from the LT forum threads that were not explicitly marked up by forum members. In addition, we added annotated example books with each topic statement, so that participants can investigate the value of query-by-example techniques in combination with more traditional text-based queries.

We found that the manually extracted suggestions increase the recall base but also further skew the distribution of suggestions, with more books receiving multiple suggestions, thereby increasing their relevance value.

The evaluation has shown that the most effective systems either adopt a learning-to-rank approach or incorporate keywords from the example books in the textual query. The effectiveness of learning-to-rank approaches suggests the complexity of dealing with multiple sources of evidence—book descriptions by

multiple authors, differing in nature from controlled vocabulary descriptors, free-text tags and full-text reviews and information needs and interests represented by both natural language statements and user profiles—requires optimizing parameters through observing users' interactions.

Next year, we continue this focus on complex topics with example books and consider including an recommender systems type evaluation. 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

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.

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.

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.

M. Koolen, T. Bogers, J. Kamps, G. Kazai, and M. Preminger. Overview of the INEX 2014 social book search track. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, volume 1180 of *CEUR Workshop Proceedings*, pages 462–479. CEUR-WS.org, 2014.

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