Overview of the CLEF 2016 Social Book Search Lab

Marijn Koolen[1,2], Toine Bogers[3], Maria Gäde[4], Mark Hall[5], Iris Hendrickx[6], Hugo Huurdeman[1], Jaap Kamps[1], Mette Skov[7], Suzan Verberne[6], and David Walsh[5]

[1] University of Amsterdam, Netherlands
{marijn.koolen,huurdeman,kamps}@uva.nl
[2] Netherlands Institute for Sound and Vision
mkoolen@beeldengeluid.nl
[3] Aalborg University Copenhagen, Denmark
toine@hum.aau.dk
[4] Humboldt University Berlin, Germany
maria.gaede@ibi.hu-berlin.de
[5] CLS/CLST, Radboud University, Nijmegen, Netherlands
(i.hendrickx|s.verberne)@let.ru.nl
[6] Edge Hill University, United Kingdom
{mark.hall,david.walsh}@edgehill.ac.uk
[7] Aalborg University, Denmark
skov@hum.aau.dk

Abstract. The Social Book Search (SBS) Lab investigates book search in scenarios where users search with more than just a query, and look for more than objective metadata. Real-world information needs are generally complex, yet almost all research focuses instead on either relatively simple search based on queries, or on profile-based recommendation. The goal is to research and develop techniques to support users in complex book search tasks. The SBS Lab has three tracks. The aim of the Suggestion Track is to develop test collections for evaluating ranking effectiveness of book retrieval and recommender systems. The aim of the Interactive Track is to develop user interfaces that support users through each stage during complex search tasks and to investigate how users exploit professional metadata and user-generated content. The Mining Track focuses on detecting and linking book titles in online book discussion forums, as well as detecting book search research in forum posts for automatic book recommendation.

1 Introduction

The goal of the Social Book Search (SBS) Lab[1] is to evaluate approaches for supporting users in searching collections of books. The SBS Lab investigates the complex nature of relevance in book search and the role of traditional and user-generated book metadata in retrieval. The aims are (1) to develop test collections

[1] See: http://social-book-search.humanities.uva.nl/

for evaluating information retrieval systems in terms of ranking search results; (2) to develop user interfaces and conduct user studies to investigate book search in scenarios with complex information needs and book descriptions that combine heterogeneous information from multiple sources; and (3) to develop algorithms that can automatically detect book search requests and suggestions from online discussions.

The SBS Lab runs three tracks:

- *Suggestion*: this is a system-centred track focused on the comparative evaluation of systems in terms of how well they rank search results for complex book search requests that consist of both extensive natural language expressions of information needs as well as example books that reflect important aspects of those information needs, using a large collection of book descriptions with both professional metadata and user-generated content.
- *Interactive*: this is a user-centred track investigating how searchers use different types of metadata at various stages in the search process and how a search interface can support each stage in that process.
- *Mining*: this is a new track focused on detecting book search requests in forum posts for automatic book recommendation, as well as detecting and linking book titles in online book discussion forums.

In this paper, we report on the setup and results of the 2016 Suggestion and Interactive Tracks as part of the SBS Lab at CLEF 2016. The three tracks run in close collaboration, all focusing on the complex nature of book search.

2 Participating Organisations

A total of 40 organisations registered for the 2016 SBS Lab, of which 29 registered for the Suggestion Track, 19 for the Interactive Track and 28 for the Mining Track. In the Suggestion Track, 10 organisations submitted runs, compared to 11 in 2015 and 8 in 2014. In the Interactive Track, 7 organisations recruited users, compared to 7 in 2015 and 4 in 2014. In the Mining Track, which ran for the first time this year, 4 organisations submitted runs. The active organisations are listed in Table 1. Participation in the SBS Lab seems to have stabilised.

3 The Amazon/LibraryThing Corpus

For all three tracks 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.[2]

[2] See http://social-book-search.humanities.uva.nl/#/collection for information on how to gain access to the corpus.

Table 1. Active participants of the CLEF 2015 Social Book Search Lab, tracks they were active in (I=Interactive, M=Mining, S=Suggestion) and number of contributed runs or users.

Institute	**Acronym**	**Tracks**	**Runs/ Users**
Aalborg University	AAU	I	14
Aix-Marseille Université CNRS	LSIS	M, S	8, 4
Chaoyang University of Technology	CYUT	S	6
Edge Hill University	Computing@EHU	I	12
Indian School of Mines Dhanbad	ISMD	S	6
Tunis EL Manar University	LIPAH	M	6
Humboldt University, Berlin	Humboldt	I	7
Know-Center	Know	M, S	8, 2
Laboratoire d'Informatique de Grenoble	MRIM	S	6
Manchester Metropolitan University	MMU	I	13
Oslo & Akershus University College of Applied Sciences	OAUC	I, S	15, 3
Peking University, China and Stockholm University, Sweden	ChiSwe	I	29
Radboud University Nijmegen	RUN	M	12
Research Center on Scientific and Technical Information	CERIST	S	6
University of Amsterdam	UvA	S	1
University of Duisburg-Essen	WGIS	I	21
University of Neuchtel, Zurich University of Applied Sciences	UniNe-ZHAW	S	6
University of Science and Technology Beijing	USTB_PRIR	S	6
Total		I, M, S	111, 32, 46

The collection consists of 2.8 million book records from Amazon including user reviews, extended with social metadata from LibraryThing (LT). This set represents the books available through Amazon. Each book is identified by an ISBN. Popular works have multiple ISBNs, so often have multiple records in the collection. Based on an ISBNs to work mapping provided by LibraryThing,[3] the 2.8 million records represent 2.4 million distinct works. 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.

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

[3] See urlhttp://www.librarything.com/feeds/thingISBN.xml.gz

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. There are 1,248,816 records from the LoC and 1,158,070 records in MARC format from the BL. Combined, these 2,406,886 records cover 1,823,998 of the ISBNs in the A/LT collection (66%).

4 Suggestion Track

4.1 Track Goals and Background

The goal of the Suggestion 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. 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 [13], 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 [7] or recommendation.

The Suggestion 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 [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. In addition, we distributed a set of 94,656 user profiles containing over 33 million book cataloguing transactions. These contain an anonymised user name, book ID, book title, author, user rating and tags and cataloguing date.

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?

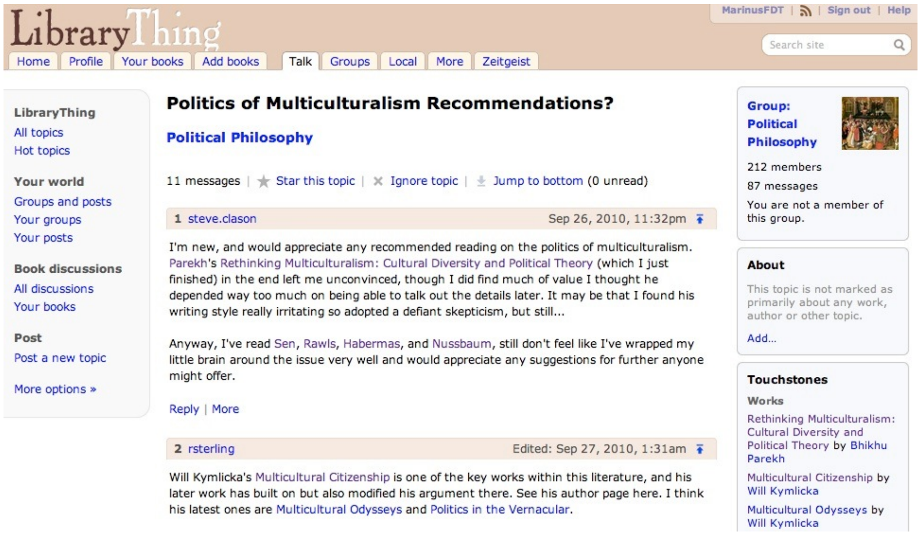
Task description The task is to reply to a user request posted on a LT forum (see Section 4.2) by returning a list of recommended books matching the user's information need. More specifically, the task assumes a user who issues a query to a retrieval system, which then returns a (ranked) list of relevant 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. 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.

4.2 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 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. We note that LT users may discuss their search requests and suggestions outside of the LT forums as well, e.g. share links of their forum request posts on Twitter. To what extent the suggestions made outside of LT differ or complement those on the forums requires investigation.

Topic selection The topic set of 2016 is a newly selected set of topics from the LibraryThing discussion forums. A total of 2000 topic threads were assessed on whether they contain a book search request by four judges, with 272 threads



Politics of Multiculturalism Recommendations?

Political Philosophy

1 steve.clason Sep 26, 2010, 11:32pm

I'm new, and would appreciate any recommended reading on the politics of multiculturalism. Parekh's Rethinking Multiculturalism: Cultural Diversity and Political Theory (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 Sen, Rawls, Habermas, and Nussbaum, 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.

2 rsterling Edited: Sep 27, 2010, 1:31am

Will Kymlicka's Multicultural Citizenship is one of the key works within this literature, and his later work has built on but also modified his argument there. See his author page here. I think his latest ones are Multicultural Odysseys and Politics in the Vernacular.

Group: Political Philosophy

212 members
87 messages
You are not a member of this group.

About

This topic is not marked as primarily about any work, author or other topic.

Add...

Touchstones

Works

Rethinking Multiculturalism: Cultural Diversity and Political Theory by Bhikhu Parekh

Multicultural Citizenship by Will Kymlicka

Multicultural Odysseys by Will Kymlicka

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

labelled as book search requests. To establish inter-annotator agreement, 453 threads were double-assessed, resulting a Cohen's Kappa of 0.83. Judges strongly agree on which posts are book search requests and which are not. Of these 272 book search requests, 124 (46%) are known-item searches from the *Name that Book* discussion group. Here, LT members start a thread to describe a book they know but cannot remember the title and author of and ask others for help. In earlier work we found that known-item topics behave very differently from the other topic types [10]. We remove these topics from the topic set so that they do not dominate the performance comparison. Furthermore, we removed topics that have no book suggestions by other LT members and topics for which we have no user profile of the topic starter, resulting in a topic set of 120 topics for evaluation of the 2016 Suggestion Track. Below is one topics in the format as it was distributed to participants.

```
<topic>
  <topicid>107277</topicid>
  <request>Greetings! I'm looking for suggestions of fantasy
  novels whose heroines are creative in some way and have some
  sort of talent in art, music, or literature. I've seen my
  share of "tough gals" who know how to swing a sword or throw a
  punch but have next to nothing in the way of imagination. I'd
  like to see a few fantasy-genre Anne Shirleys or Jo Marches.

  Juliet Marillier is one of my favorite authors because she
  makes a point of giving most of her heroines creative talents.
  Even her most "ordinary" heroines have imagination and use it
  to create. Clodagh from "Heir to Sevenwaters," for example,
  may see herself as being purely domestic, but she plays the
  harp and can even compose songs and stories. Creidhe of
  "Foxmask" can’t read, but she can weave stories and make
  colors. The less ordinary heroines, like Sorcha from "Daughter
  of the Forest" and Liadan from "Son of the Shadows," are good
  storytellers. I’m looking for more heroines like these.

  Any suggestions?</request>
  <group>FantasyFans</group>
  <title>Fantasy books with creative heroines?</title>
  <examples>
    <work>
      <booktitle>Daughter of the Forest</booktitle>
      <author>Juliet Marillier</author>
      <workid>6442</workid>
    </work>
    <work>
      <booktitle>Foxmask</booktitle>
      <author>Juliet Marillier</author>
      <workid>349475</workid>
    </work>
    <work>
      <booktitle>Son of the Shadows</booktitle>
      <author>Juliet Marillier</author>
      <workid>6471</workid>
    </work>
  </examples>
  <catalogue>
    <work>
      <tags/>
      <rating>0.0</rating>
      <publication-year>2002</publication-year>
      <booktitle>Blue Moon (Anita Blake, Vampire Hunter, Book 8)</booktitle>
      <cataloging-date>2011-08</cataloging-date>
      <author>Laurell K. Hamilton</author>
      <workid>10868</workid>
    </work>
    ...
</catalogue>
 </topic>
```

Table 2. 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
USTB-PRIR	run1.keyQuery_active_combineRerank	0.2157	0.5247	0.1253	0.3474
CERIST	all_features	0.1567	0.3513	0.0838	0.4330
CYUT-CSIE	0.95Averageword2vecType2TGR	0.1158	0.2563	0.0563	0.1603
UvA-ILLC	base_es	0.0944	0.2272	0.0548	0.3122
MRIM	RUN2	0.0889	0.1889	0.0518	0.3491
ISMD	ISMD16allfieds	0.0765	0.1722	0.0342	0.2157
UniNe-ZHAW	Pages_INEXSBS2016_SUM_SCORE	0.0674	0.1512	0.0472	0.2556
LSIS	Run1_ExeOrNarrativeNSW_Collection	0.0450	0.1166	0.0251	0.2050
OAU	oauc_reranked_ownQueryModel	0.0228	0.0766	0.0127	0.1265
know	sbs16suggestiontopicsresult2	0.0058	0.0227	0.0010	0.0013

Operationalisation of forum judgement labels In previous years, the Suggestion Track used a complicated decision tree to derive a relevance value from a suggestion. To reduce the number of assumptions, we simplified the mapping of book suggestions to relevance values. By default a suggested book has a relevance value of 1. Books that the requester already has in her personal catalogue before starting the thread (pre-catalogued suggestions) have little additional value and are assumed to have a relevance value of 0. On the other hand, suggestions that the requester subsequently adds to her catalogue (post-catalogued suggestions) are assumed to be the most relevant suggestions and receive a relevance value of 8, to keep the relevance level the same as in 2014 and 2015. Note that some of the books mentioned in the forums are not part of the 2.8 million books in our collection. We therefore removed any books from the suggestions that are not in the INEX A/LT collection. The numbers reported in the previous section were calculated after this filtering step.

4.3 Evaluation

This year, 10 teams submitted a total of 46 runs (see Table 1). The evaluation results are shown in Table 2 for the best run per team. 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 are also reported.

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

1. *USTB-PRIR - run1.keyQuery_active_combineRerank* (rank 1): This run was made by a searching-re-ranking process where the initial retrieval result was based on the selection of query keywords and a small index of active books, the re-ranking results based on a combination of several strategies (number of people who read the book from profile, similar-product from amazon.com, popularity from LT forum, etc.).
2. *CERIST - all_features* (rank 7): The topic statement in the *request* field is treated as a verbose query and is reduced using several features based on term statistics, Part-Of-Speech tagging, and whether terms from the *request* field occur in the user profile and example books.

3. *CYUT-CSIE - 0.95Averageword2vecType2TGR* (rank 11): This run uses query expansion based on word embeddings using word2vec, on top of a standard Lucene index and retrieval model. For this run, queries are represented by a combination of the *title*, *group* and *request* fields. Results are re-ranked using a linear combination of the original retrieval score and the average Amazon user ratings of the retrieved books.
4. *UvA-ILLC - base_es* (rank 17): This run is based on a full-text ElasticSearch [3] index of the A/LT collection, where the Dewey Decimal Codes are replaced by their textual representation. Default retrieval parameters are used, the query is a combination of the topic *title*, *group* and *request* fields. This is the same index that is used for the experimental system of the Interactive Track (see Section 5.3) and serves as a baseline for the Suggestion Track.
5. *MRIM - RUN2* (rank 18): This run is a weighted linear fusion of a BM25F run on all fields, an Language Model (LM) run on all fields, and two query expansion runs, based on the BM25 and LM run respectively, using as expansion terms an intersection of terms in the user profiles and word embeddings from the query terms.

Most of the top performing systems, including the top performing run preprocess the rich topic statement with the aim of reducing the *request* to a set of most relevant terms. Two of the top five teams use the user profiles to modify the topic statement. This is the first year that word embeddings are used for the Suggestion Track. Both CYUT-CSIE and MRIM found that word embeddings improved performance over configurations without them. From these results it seems clear that topic representation is an important aspect in social book search. The longer narrative of the *request* field as well as the metadata in the user profiles and example books contain important information regarding the information need, but many terms are noisy, so a filtering step is essential to focus on the user's specific needs.

5 Interactive Track

The goal of the Interactive Track is to investigate how searchers make use of and combine professional metadata and user-generated content for book search on the Web and to develop interfaces that support searchers through the various stages of their search task. Through user-generated content, book descriptions are extended far beyond what is traditionally stored in professional catalogues. Not only are books described in the users' own vocabulary, but they are also reviewed and discussed online. As described in Section 4, this subjective user-generated content can help users during search tasks where their personal preferences, interests and background knowledge play a role.

The Interactive 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 uses a subset of 1.5

million out of 2.8 million records of the A/LT collection (described in Section 3) for which a thumbnail cover image is available.

5.1 User Tasks

Participants started with a training task to ensure that they were familiar with system's functionality. Next, participants were asked to complete one mandatory task which was either a *goal-oriented* task (56 participants) or a *non-goal* task (55 participants). After completing the mandatory task participants were asked whether they had time to complete an *optional* task. 89 participants completed one of the eight optional tasks.

The *goal-oriented* task contains five sub-tasks ensuring that participants spend enough time to generate a rich data-set. While the first sub-task defines a clear goal, the other sub-tasks are more open to give participants the flexibility to interact with the available content and functionality. The same instruction text was used as in the 2015 track [8].

The *non-goal* task was developed based on the open-ended task used in the iCHiC task at CLEF 2013 [14] and the ISBS task at CLEF 2014 [6]. The aim of this task is to investigate how users interact with the system when they have no pre-defined goal in a more exploratory search context. It also allows the participants to bring their own goals or sub-tasks to the experiment in line with the "simulated work task" idea [2]. The same instruction text was used as in the 2015 track [8].

The *optional* tasks represent real Library Thing forum requests. 89 participants indicated that they had time for an optional task and were randomly given one of eight optional tasks, that were selected from the tasks used in the suggestion track. An example of an optional task:

> You're interested in non-fiction history books on the background to and the actual time of the Boer War in South Africa. Search the collection using any of the interface features to find at least one book that meets these criteria.

5.2 Experiment Structure

The experiment was conducted using the SPIRE system[4] [4], using the flow shown in Figure 2. When a new participant started the experiment, the SPIRE system automatically allocated them either the *non-goal* or *goal-oriented*task. If they chose to undertake the *optional* task, they were also allocated one of the

[4] Based on the Experiment Support System – https://bitbucket.org/mhall/experiment-support-system

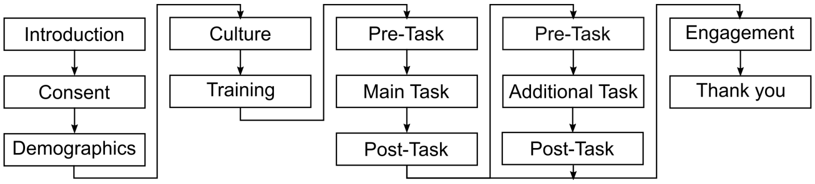


Fig. 2. The path participants took through the experiment. The SPIRE system automatically balanced task allocation in both the *Main Task* and *Additional Task*. After the first *Post-Task* stage, participants were asked whether they had time to do another task and if not, were taken directly to the *Engagement* stage.

eight *optional* tasks. The SPIRE system automatically balances task allocation for both the main and optional tasks. Additionally each participating team was allocated their own experiment instance to ensure optimal balance both within and across the teams. Participants were not explicitly instructed to use only the interface and collection provided, so it is possible some users used other websites as well. However, given the lack of incentive to use external websites, we expect this issue to be negligible.

Participant responses were collected in the following five steps using a selection of standard questionnaires:

- *Consent* – participants had to confirm that they understood the tasks and the types of data collected in the experiment.
- *Demographics* – gender, age, achieved education level, current education level, and employment status;
- *Culture* – country of birth, country of residence, mother tongue, primary language spoken at home, languages used to search the web;
- *Post-Task* – after each task, participants judged the usefulness of interface components and meta-data parts, using 5-point Likert-like scales;
- *Engagement* – after completing both tasks, they were asked to complete O'Brien et al.'s [12] engagement scale.

5.3 System and Interfaces

The user interface was both built using the PyIRE[5] workbench, which provides the required functionality for creating interactive IR interfaces and logging all interactions between the participants and the system. This includes any queries they enter, the books shown for the queries, pagination, facets selected, books viewed in detail, metadata facets viewed, books added to the book-bag, and books removed from the book-bag. All log-data is automatically timestamped and linked to the participant and task.

[5] Python interactive Information Retrieval Evaluation workbench – https://bitbucket.org/mhall/pyire

The backend IR system was implemented using ElasticSearch[6], which provided free-text search, faceted search, and access to the individual books' complete metadata. This is index was also used as a baseline system in the 2016 Suggestion Track (see Section 4.3). The 1.5 million book descriptions are indexed with all professional metadata and user-generated content. For indexing and retrieval the default parameters are used, which means stopwords are removed, but no stemming is performed. The Dewey Decimal Classification numbers are replaced by their natural language description. That is, the DDC number 573 is replaced by the descriptor *Physical anthropology*. User tags from LibraryThing are indexed both as text strings, such that complex terms are broken down into individual terms (e.g. *physical anthropology* is indexed as *physical* and *anthropology*) and as non-analyzed terms, which leaves complex terms intact and is used for faceted search.

The interface was designed to support users by taking the different stages of the search process into account. The idea behind the *multi-stage* interface design was inspired by Kuhlthau [11] and Vakkari [15] and it includes three distinct panels, potentially supporting different stages: *browse*, in which users can explore categories of books, *search*, supporting in-depth searching, and *book-bag*, in which users can review and refine their book-bag selections. An earlier model of decision stages in book selection [13] supports the need for a user interface that takes the different search and decision stages into account.

When the *multi-stage* interface first loads, participants are shown the *browse* stage, which is aimed at supporting the initial exploration of the data-set. The main feature to support the free exploration is the hierarchy browsing component on the left, which shows a hierarchical tree of Amazon subject classifications. This was generated using the algorithm described in [5], which uses the relative frequencies of the subjects to arrange them into the tree-structure with the most-frequent subjects at the top of the tree. The search result list is designed to be more compact to allow the user to browse books quickly and shows only the book's title, thumbnail image, and aggregate ratings (if available). Clicking on the book title showed a popup window with the book's full meta-data.

Participants switched to the *search* stage by clicking on the "Search" section in the gray bar at the top. The *search* stage presents a standard faceted search interface. Additionally if the participant had selected a topic in the *explore* stage, then the search was initially filtered by this as well. Participants could then select to search the whole collection.

The final stage is the *book-bag*, where participants review the books they have collected and can provide notes for each book. For each book participants could search for similar books by title, author, topic, and user tags, using the same compact layout as in the *browse* stage.

[6] ElasticSearch – http://www.elasticsearch.org/

5.4 Participants

A total of 111 participants were recruited (see Table 1), 51 female and 60 male. 65 were between 18 and 25, 29 between 26 and 35, 16 between 36 and 45, and 1 between 46 and 55. 31 were in employment, 2 unemployed, 77 were students and 1 selected *other*. Participants came from 15 different countries (country of birth) including China (27 participants), UK (25), Norway (14), Germany (13), India (11), Denmark (10), resident in 8 different countries, again mainly in China, UK, Germany, Norway and Denmark. Participants' mother tongues included Chinese, English, German, Norwegian and 9 others. The majority of participants executed the tasks in a lab (74) and only 37 users participated remotely.

5.5 Procedure

The participants were invited by the individual teams, as described in Section 5.2, either using e-mail or by recruiting students from a lecture or lab. The following browsers and operating systems had been tested: Windows, OS X, Linux using Internet Explorer, Chrome, Mozilla Firefox, and Safari. The only difference between browsers was that some of the graphical refinements such as shadows are not supported on Internet Explorer and fall back to a simpler line-based display.

After participants had completed the experiment as outlined above (5.2), they were provided with additional information on the tasks they had completed and with contact information, should they wish to learn more about the experiment. Where participants that completed the experiment in a lab, teams were able to conduct their own post-experiment process.

5.6 Results

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

Session length was measured using JavaScript. Table 3 shows median and inter-quartile ranges for all tasks. The data show clear distinctions between *non-goal*, *goal-oriented*, and optional tasks.

Number of queries was extracted from the log-data. Queries could be issued by typing keywords into the search box or by clicking on a meta-data field to search for other books with that meta-data field value. Both types of query have been aggregated and Table 3 shows the number of queries for each task. There is a clear difference between the *non-goal* and the *goal-oriented* task. On the *additional* tasks, more analysis is needed to investigate why the *south africa*, *complex mystery*, and *romance mystery* tasks have such low values for the number of queries. However, for the other *additional* tasks, it is clear that as far as complexity of the task and number of queries required, they lie between the *non-goal* and *goal-oriented* tasks.

Table 3. Session lengths for the tasks. Times are in minutes:seconds and are reported median (inter-quartile range); Queries and Books Collected are reported median (inter-quartile range).

Task	Time	# Queries	# Books
Non-goal	7:38min (9:38min)	1 (5)	3 (3)
Goal-oriented	12:20min (14:28min)	5 (9)	5 (0)
South Africa	4:51min (3:51min)	1 (1.5)	2.5 (2)
Elizabethan	5:48min (3min)	3.5 (3.25)	2.25 (2.25)
Communication	4:58min (2:47min)	4 (4)	2 (2)
Painters	4:42min (4:07min)	4 (4)	1 (2)
Complex Mystery	3:36min (4:22min)	0 (0)	1 (0)
Astronomy	5:12min (5:54min)	3 (3.25)	2 (1)
Romance Mystery	2:21min (3:15min)	1 (2)	2 (1)
French Revolution	6:36min (7:16min)	4 (4.5)	1 (0)

Number of books collected was extracted from the log-data. The numbers reported in Table 3 are based on the number of books participants had in their book-bag when they completed the session, as participants could remove books they had previously collected.

The number of books collected is determined by the task, although the *elizabethan* and *south africa*, and *communication* tasks have different potential interpretations on how many books are needed to satisfy the task. As is to be expected, the *non-goal* task shows the highest variation in the number of books collected, as participants were completely free to define what "success" meant for them in that task.

6 Mining Track

6.1 Track Goals and Background

The Mining track is a new addition to the Social Book Search Lab in 2016. The goal of the Mining Track is twofold: (1) to detect book search requests in forum posts for automatic book recommendation, and (2) to detect and link book titles in online book discussion forums. The mining track represents the first stage in supporting complex book-related information needs. Later stages, such as the retrieval stage and user interaction with book search engines, have already been investigated in the Suggestion and Interactive tracks.

Up to now, these tracks have relied on the manual identification, analysis, and classification of complex search tasks as expressed in the LT discussion fora to serve as input in these tracks, as described in Section 4. Book search requests were manually separated from other book-related discussion threads by human annotators, and the suggestions provided by other LT users were used as relevance judgments in the automatic evaluation of retrieval algorithms that were applied to the book search requests.

However, to be able to fully support complex book search behavior, we should not just support the (interactive) retrieval and recommendation stage of the process, but also the automatic detection of complex search needs and the analysis of these needs and the books and authors contained therein. This is the goal of the Mining Track. The first edition of the Mining Track focuses on automating two text mining tasks in particular:

1. **Book search request classification**, in which the goal is to identify which threads on online forums are book search requests. That is, given a forum thread, the system should determine whether the opening post contains a request for book suggestions (i.e., binary classification of opening posts)
2. **Book linking**, in which the goal is to recognize book titles in forum posts and link them to the corresponding metadata record through their unique book ID. The task is not to mark each entity mention in the post text, but to label the post as a whole with the IDs of the mentioned books. That is, the system does not have to identify the exact phrase that refers to book, but only has to identify which book is mentioned on a per-post basis.

6.2 Track setup

Task 1: Classifying forum threads

Data collection For the task of classifying forum threads we created two data sets for training: one based on the LT forums and one based on Reddit. For the LT forums, we randomly sampled 4,000 threads and extracted their opening posts. We split them into a training and a test set, each containing 2,000 threads. These threads contained both positive and negative examples of book requests.

The Reddit training data was sampled from three months of Reddit opening posts published in September, October, and November 2014. The set of positive book request examples comprises all threads from the suggestmeabook subReddit, whereas the negative examples comprises all threads from the books subReddit. The training set contained 248 threads in total with 43 positive and 205 negative examples. The Reddit test data was sampled from December 2014 and comprises 89 threads with 76 negative and 13 positive examples of book requests.

Annotation The labels of the Reddit *training* data were not annotated manually, as they were already categorized as positive and negative by virtue of the subReddit (books or suggestmeabook) they originated from. The Reddit *test* set originally consisted of 89 threads with the subReddit names as labels. In order to create a reliable ground truth for the test set, two track organizers manually classified the 89 test threads. All disagreements were discussed and we reached consensus on all 89 threads. 81 of the labels were the same as the original Reddit label; the other 8 were different. We use the manual labels as ground truth.

In the annotation process for the LT threads, positive examples of book requests consisted of all posts where the user described an explicit foreground or background information need and was searching for books to read. Examples include *known-item* requests, where a user is looking for a specific book by describing plot elements, but cannot remember the title; users asking for books covering a specific topic; and users asking for books that are similar to another book they mention. Posts where users ask for new authors to explore or where they list their favorite books and ask others to do the same are *not* classified as explicit book requests.

Task 2: Book linking

Data collection Book linking through the use of so-called 'touchstones' is an striking characteristic of the LT forum, and an important feature for the forum community. A touchstone is a link created by a forum member between a book mention in a forum post and a unique LT work ID in the LT database. A single post can have zero or more different touchstones linked to it. Touchstones allow readers of a forum thread to quickly see which books are mentioned in the thread.

For the book linking task we created a data set based on the touchstones in the LT forum. The training data consisted of 200 threads with 3619 posts in total. The test data for the linking task comprised 200 LT threads with 3809 posts in total. The task is to identify the LT work ID of all unique books mentioned in the post and link them to their specific post IDs.

In addition to the training data set, participants were encouraged to use the Amazon/LT collection used in the Suggestion Track to aid in the book linking task. This collection contains 2.8 million book metadata records along with their LT work IDs.

Annotation In the annotation process, we annotated the posts in the LT threads (up to a maximum of 50 posts per thread) with all touchstones that were not added by LT users yet. Preliminary analysis has shown that around 16% of all books are not linked [9]. We manually linked book titles in the posts to their unique LT work ID. Many books are published in different editions throughout the years with different unique ISBNs, but all of these versions are connected to the same unique LT work ID. If a book occurred multiple times in the same post, only the first occurrence was linked, so participants only need to specify each of the work IDs found in a post once. If a post mentioned a series of books, we linked this series to the first book in the series, e.g., the Harry Potter series was linked to "Harry Potter and the Philosopher's Stone". We did not link book authors. In some cases, a book title was mentioned, but no suitable work ID was found in the Amazon/LT collection. In this cases, we labeled that book title as UNKNOWN.

The annotation of book titles was found to be a difficult task for several reasons, in particular: (a) the definition of 'book reference' is not trivial: all sorts of abbreviations and author references are made; and (b) finding the book that is referred to, is sometimes difficult due to ambiguities and errors in titles. The latter was even more challenging in the case of book series.

In total, the dataset of 200 threads comprises 5097 book titles in 2117 posts.

6.3 Evaluation

For the book request classification task, we computed and report only accuracy, because this is a single-label, binary classification task. For the linking task, we computed accuracy, precision, recall, and F-score.

Both tasks are performed and evaluated at the level of forum posts. We detect whether a forum post was a book request in the classification task, and whether a certain book title occurred in a post. In case the same book title was mentioned multiple times in the same post, we only count and evaluate one occurrence of this particular book title. Each book title was mapped to a LT work ID that links together different editions of the same book (with different ISBNs).

During manual annotation, we came across several book titles for which we were unable to find the correct LT work ID. These cases were problematic in the evaluation: just because the annotator could not find the correct work ID, that does not mean it does not exist. For that reason, we decided to discard these examples in the evaluation of the test set results. In total, 180 out of the 5097 book titles in the test set were discarded for this reason.

Similarly, during the book request classification task, we also found some cases where we were unsure about categorizing them as book search requests or not and we discarded 26 such cases from the test set in the evaluation.

6.4 Results

Task 1: Classifying forum threads

For evaluation, 1,974 out of the 2,000 threads in the LT test set were used. For the 26 remaining threads, judges were unsure whether the first post was a request or not.

For the baseline system of the classification task, we trained separate classifiers for the two data sets (LT and Reddit) using scikit-learn[7]. We extracted bag-of-words-features (either words or character 4-grams) from the title and the body of the first post, and for LT also from the category (for Reddit, the category was the label). We used tf-idf weights for the words and the character 4-grams from these fields. We ran 3 classifiers on these data: Multinonial Naive Bayes (MNB), Linear Support Vector Classification (LinearSVC) and KNN, all with their default hyperparameter settings in scikit-learn.

Table 4 shows the results on the book search request classification task. We observe a clear difference in performance of the system on the LT and Reddit test sets.

Task 2: Book linking

For evaluation, 217 out of the 220 threads in the test set were used, with 5097 book titles identified in 2117 posts. A further 180 book titles were found

[7] http://scikit-learn.org/

Table 4. Results for the classification task for the two datasets in terms of accuracy on the 1974 LT and 89 Reddit posts.

		LT		Reddit	
Team	**Run**	**Rank**	**Accuracy**	**Rank**	**Accuracy**
baseline	character_4-grams.LinearSVC	1	94.17	6	78.65
baseline	Words.LinearSVC	2	93.92	5	78.65
Know	Classification-Naive-Resutls	3	91.59	2	82.02
baseline	character_4-grams.KNeighborsClassifier	4	91.54	7	78.65
baseline	Words.KNeighborsClassifier	5	91.39	4	78.65
LIPAH	submission2-librarything	6	90.98	-	-
LIPAH	submission3-librarything	7	90.93	-	-
LIPAH	submission4-librarything	8	90.83	-	-
Know	Classification-Veto-Resutls	9	90.63	9	76.40
LIPAH	submission1-librarything	10	90.53	-	-
baseline	character_4-grams.MultinomialNB	11	87.59	11	76.40
baseline	Words.MultinomialNB	12	87.59	10	76.40
Know	Classification-Tree-Resutls	13	83.38	8	76.40
Know	Classification-Forest-Resutls	14	74.82	12	74.16
LIPAH	submission6-Reddit	-	-	1	82.02
LIPAH	submission5-Reddit	-	-	3	80.90

that could not be linked to works in the book metadata set of 1,925,024 books. These 180 unlinked book titles are ignored in the evaluation. Table 5 shows the results on the book linking task.

Table 5. Results for the linking task for the LT data set in terms of accuracy

Rank	Team	Run	# posts	Accuracy	Recall	Precision	F-score
1	Know	sbs16classificationlinking	4917	41.14	41.14	28.26	33.50
2	LSIS	BA_V2bis	4917	26.99	26.99	38.23	31.64
3	LSIS	BA_V1bis	4917	26.54	26.54	37.58	31.11
4	LSIS	B_V2bis	4917	26.01	26.01	35.39	29.98
5	LSIS	BUbis	4917	26.34	26.34	34.50	29.87
6	LSIS	Bbis	4917	25.54	25.54	34.80	29.46

7 Conclusions and Plans

This was the second year of the Social Book Search (SBS) Lab. The SBS Lab investigates book search in social media from three different perspectives: 1) the evaluation of retrieval and ranking algorithms for complex book search tasks in the Suggestion Track, 2) studying how systems can support users in different phases of these complex search tasks in the Interactive Track, and 3) evaluating how well systems can identify book search tasks and book mentions in online book discussions in the Mining Track.

The Suggestion Track was changed little from the previous edition in 2015. In selecting topics, known-item information needs were removed to focus on recommendation requests. The user profiles and topic representations were enriched with more book metadata compared to previous years to give more information about the users and their information needs. Several of the best performing systems incorporated techniques for summarizing and reducing the rich natural language topic statement to remove irrelevant terms and focus on the need. Word embeddings were successfully used by several participants both for expanding queries and for summarizing the topic statements.

For the Interactive Track we simplified the experimental setup compared to 2015, such that users did only one mandatory task with at most one optional task. The optional tasks were based on book search requests from the LT forums, which result in notably different behaviour from the artificially created goal-oriented task.

The Mining Track ran for the first time in 2016, with the aim of evaluating systems that automatically identify book search requests and book mentions in online book discussions. Typical for the first year of a task, several issues with the task and its evaluation were identified. The classification task appeared to be straightforward, both in annotation as in implementation. The results show that the task is feasible and can be performed automatically with a high accuracy. The book linking task however posed a number of challenges, especially in annotating the data. The small number of participants in the track does not allow us to make informative comparisons between multiple different approaches to the tasks.

The 2016 SBS Lab saw the introduction of one new track and created further ways in which the tracks can collaborate and mutually inform each other. For the 2017 Interactive, we plan to introduce new features in the multistage system, such as building shortlists for searching with multiple example books and comparing the metadata of shortlisted books to build richer query representations. We expect these features will allow us to further investigate search stages and search strategies. The optional tasks in the 2016 Interactive Track have given us rich user interactions for a number of real-world complex book search information needs which we plan to use in the Suggestion Track as more structured representations of the information need. For the Mining Track the next step would be to expand and improve the two mining tasks in order to embed them in the social book search pipeline: starting with a complex book search request, find book titles that are relevant book suggestions and link them to their unique identifier. Alternatively, the classification task could be expanded to include the classification of different types of book search requests.

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. P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of documentation*, 53(3):225–250, 1997.
3. Elastic. Elasticsearch, 2016. URL https://www.elastic.co/products/elasticsearch.
4. M. M. Hall and E. Toms. Building a common framework for iir evaluation. In *CLEF 2013 - Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 17–28, 2013. doi: 10.1007/978-3-642-40802-1_3.
5. M. M. Hall, S. Fernando, P. Clough, A. Soroa, E. Agirre, and M. Stevenson. Evaluating hierarchical organisation structures for exploring digital libraries. *Information Retrieval*, 17(4):351–379, 2014. ISSN 1386-4564. doi: 10.1007/s10791-014-9242-y. URL http://dx.doi.org/10.1007/s10791-014-9242-y.
6. M. M. Hall, H. C. Huurdeman, M. Koolen, M. Skov, and D. Walsh. Overview of the INEX 2014 interactive social book search track. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, volume 1180 of *CEUR Workshop Proceedings*, pages 480–493. CEUR-WS.org, 2014. URL http://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-HallEt2014.pdf.
7. 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.
8. M. Koolen, T. Bogers, M. Gäde, M. Hall, H. Huurdeman, J. Kamps, M. Skov, E. Toms, and D. Walsh. *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings*, chapter Overview of the CLEF 2015 Social Book Search Lab, pages 545–564. Springer International Publishing, Cham, 2015. ISBN 978-3-319-24027-5. doi: 10.1007/978-3-319-24027-5_51. URL http://dx.doi.org/10.1007/978-3-319-24027-5_51.
9. M. Koolen, T. Bogers, M. Gäde, M. A. Hall, H. C. Huurdeman, J. Kamps, M. Skov, E. Toms, and D. Walsh. Overview of the CLEF 2015 social book search lab. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. J. F. Jones, E. SanJuan, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, volume 9283 of *Lecture Notes in Computer Science*, pages 545–564. Springer, 2015.
10. M. Koolen, T. Bogers, A. van den Bosch, and J. Kamps. Looking for books in social media: An analysis of complex search requests. In A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr, editors, *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna,*

Austria, March 29 - April 2, 2015. Proceedings, volume 9022 of *Lecture Notes in Computer Science*, pages 184–196, 2015. ISBN 978-3-319-16353-6. doi: 10.1007/978-3-319-16354-3_19. URL http://dx.doi.org/10.1007/978-3-319-16354-3_19.

11. C. C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991. ISSN 1097-4571. doi: 10.1002/(SICI)1097-4571(199106)42:5⟨361::AID-ASI6⟩3.0.CO;2-#. URL http://dx.doi.org/10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-#.
12. H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2009.
13. K. Reuter. Assessing aesthetic relevance: Children's book selection in a digital library. *JASIST*, 58(12):1745–1763, 2007.
14. E. Toms and M. M. Hall. The chic interactive task (chici) at clef2013. http://www.clef-initiative.eu/documents/71612/1713e643-27c3-4d76-9a6f-926cdb1db0f4, 2013.
15. P. Vakkari. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of documentation*, 57(1):44–60, 2001.