Overview of the CLEF 2015 Social Book Search Lab

Marijn Koolen[1(✉)], Toine Bogers[2], Maria Gäde[3], Mark Hall[4], Hugo Huurdeman[1], Jaap Kamps[1], Mette Skov[5], Elaine Toms[6], and David Walsh[4]

[1] University of Amsterdam, Amsterdam, Netherlands
{marijn.koolen,h.c.huurdeman,kamps}@uva.nl
[2] Aalborg University Copehagen, Copenhagen, Denmark
toine@hum.aau.dk
[3] Humboldt University Berlin, Berlin, Germany
maria.gaede@ibi.hu-berlin.de
[4] Edge Hill University, Ormskirk, UK
{mark.hall,david.walsh}@edgehill.ac.uk
[5] Aalborg University, Aalborg, Denmark
skov@hum.aau.dk
[6] University of Sheffield, Sheffield, UK
e.toms@sheffield.ac.uk

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 recommendation based on profiles. The goal is to research and develop techniques to support users in complex book search tasks. The SBS Lab has two 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.

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 for evaluating systems in terms of ranking search results and 2) to develop user interfaces and conduct user studies to investigate book search in scenarios with complex information need and book descriptions that combine heterogeneous information from multiple sources.

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

J. Mothe et al. (Eds.): CLEF 2015, LNCS 9283, pp. 545–564, 2015.
DOI: 10.1007/978-3-319-24027-5_51

The SBS Lab runs two 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.

In this paper, we report on the setup and results of the 2015 Suggestion and Interactive Tracks as part of the SBS Lab at CLEF 2015. The two tracks run in close collaboration as both focus on the complex nature of book search. The paper is structured as follows. First, in Section 2, we give a brief summary of the participating organisations. Then, in Section 3 we provide details on the Amazon/LibraryThing corpus of book descriptions that is used for both tracks. The setup and results of the Suggestion Track are described in Section 4, followed by the experiments and results of the Interactive Track in Sections 5. We close in Section 6 with a discussion of the overall findings and plans for 2016.

2 Participating Organisations

A total of 35 organisations registered for the SBS Lab, of which 27 registered for the Suggestion Track and 28 for the Interactive Track. In the Suggestion Track, 11 organisations submitted runs, compared to 8 in 2014. In the Interactive Track, 7 organisations recruited users, compared to 4 in 2014. The active organisations are listed in Table 1.

3 The Amazon/LibraryThing Corpus

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]

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. 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

[2] See https://inex.mmci.uni-saarland.de/data/nd-agreements.jsp for information on how to gain access to the corpus.

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
Institute		# users
Aalborg University	AAU	36
University of Amsterdam	UvA	22
Edge Hill University	Edge_Hill	20
Humboldt University	Humboldt	67
Manchester Metropolitan University	MMU	23
Oslo & Akershus University College	Oslo_SBS	20
Stockholm University	Stockholm	1
Other		3
Total		192

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 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, there are 2,406,886 records covering 1,823,998 of the ISBNs in the A/LT collection (66%).

4 The SBS Suggestion Track

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 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 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 and 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 [1] described in the previous section.

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 they are also reviewed and discussed online, and added to online personal catalogues of individual readers. This additional information is subjective and personal, and opens up opportunities to aid users in searching for books in different ways that go beyond the traditional editorial metadata based search scenarios, such as known-item and subject search. For example, readers use many more aspects of books to help them decide which book to read next [9], 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 [6] or recommendation.

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

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 use 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 metadata for book search?

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. 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.

4.1 Information Needs

LT users discuss their books on the discussion forums. Many of the topic threads are started with a request from a member for interesting, fun new books to read. Users typically describe what they are looking for, give examples of what they like and do not like, indicate which books they already know and ask other members for recommendations. Members often reply with links to works catalogued on LT, which, in turn, have direct links to the corresponding records on Amazon. These requests for recommendations are natural expressions of information needs for a large collection of online book records. We use a sample of these forum topics to evaluate systems participating in the 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

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

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. Additional relevance information can be gleaned from the discussions on the threads. Consider, for example, topic 129939[4]. 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.

Topic Selection. The topic set of 2015 is a subset of the 2014 topic set, focusing on topics where the requester gives both a narrative description of the information need and one or more example books to guide the suggestions. The 2015 topic set has 208 topics, where the narrative and examples are combined with all the books of the topic creators' profiles up to the time of posting the request on the forum. This topic set was 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, 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

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

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 one the requester is neutral about. It has positive and negative aspects. From the request, forum members can understand how to interpret this example.

We had 8 annotators label each example provided by the requester and each suggestion provided by LT members. 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 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 book 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 users 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. Some of the books mentioned in the forums are not part of the 2.8 million books in our collection. We removed from the suggestions any books that are not in the 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 based on having read the book to be of more value to the searcher. For the mapping to relevance values, we refer to the first mention of work as the *suggestion* and subsequent mentions of the same work as *replies*. We use *has read* when the forum members have read the book they mention and *not read* when they have not. Furthermore, we use a number of simplifying assumptions:

- When the annotator was *not sure* if the person mentioning a book has read it, we treat it as *not read*. We argue that for the topic starter there is no clear difference in the value of such recommendations.
- When the annotator was *not sure* if a suggestion was positive, negative or neutral, we treat it as *neutral*. Again, for the topic starter there is no clear signal that there is difference in value.

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

This leads to the following graded relevance values:

- $rv = 0$: not relevant
- $rv = 1$: relevant but more negative than positive mentions
- $rv = 2$: neutral mention
- $rv = 3$: positive mention (but not read by suggester(s))
- $rv = 4$: positive mention (but not read by suggester(s))
- $rv = 6$: positive mention (read by suggester(s))
- $rv = 8$: suggestion that is afterwards catalogued by requester

More details about this mapping are provided on the Track website.[5]

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. In total, over 94,000 user profiles with 34 million cataloguing transactions were scraped from the LT site, anonymised and made available to participants. To anonymise all user profiles, we 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.

4.2 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 2.

[5] See: http://social-book-search.humanities.uva.nl/#/results15

Table 2. Evaluation results for the official submissions. Shown are the topic scoring runs for each participating team.

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	USTB_PRIR	run5-Rerank-RF-example	0.106	0.232	0.068	0.365	no
4	MRIM	LIG_3	0.098	0.189	0.069	0.514	yes
5	LaHC_Saint-Etienne	UJM_2	0.088	0.174	0.065	0.483	no
6	AAU	allfields-jm	0.087	0.191	0.061	0.420	yes
7	Oslo_SBS_iTrack_group	baseLine	0.082	0.182	0.052	0.341	no
8	CSIE	0.95AverageType2QTGN	0.082	0.194	0.050	0.319	no
9	LSIS-OpenEdition	INL2_SDM_Graph_LSIS	0.081	0.183	0.058	0.401	no
10	UAmsterdam	UAmsQTG_KNN_L.070	0.068	0.160	0.051	0.388	yes
11	IR@JU	KASIT_1	0.011	0.023	0.006	0.009	no

The best run 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 topics have been removed.
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.

5 The SBS Interactive Track

The goal of the interactive Social Book Search (ISBS) track is to investigate how searchers make use of and appreciate 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. The user has a specific information need against a background of personal tastes, interests and previously seen books. Through social media, 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. User reviews can contain information on how engaging, fun, educational or well-written a book is.

The ISBS 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, described in Section 3, using a subset of 1.5 million of the total 2.8 million book descriptions for which a thumbnail cover image is available.

5.1 User Tasks

This year in addition to the two main user tasks, a training task was developed to ensure that participants are familiar with all the functions offered by the two interfaces. The queries and topics used in the training task were chosen so as not to overlap with the *goal-oriented* task. However, a potential influence on the *non-goal* task cannot be ruled out.

Similar to last year, two tasks were created to investigate the impact of different task types on the participants interactions with the interfaces and the professional and user-generated book metadata. For both tasks, participants were asked to motivate each book selection in the book-bag.

The *Goal-Oriented* Task. contains five sub-tasks ensuring that participants spend enough time on finding relevant books. While the first sub-tasks defines a clear goal, the other sub-tasks are more open giving the user enough room to interact with and the available content and met-data options. The following instruction text was provided to participants:

> Imagine you participate in an experiment at a desert-island for one month. There will be no people, no TV, radio or other distraction. The only things you are allowed to take with you are 5 books. Please search for and add 5 books to your book-bag that you would want to read during your stay at the desert-island:

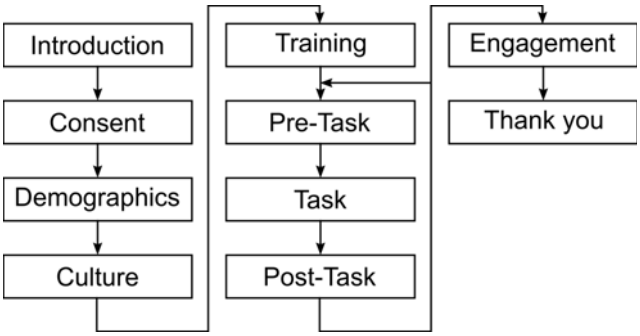

Fig. 2. The path participants took through the experiment. Each participant completed the *Pre-Task*, *Task*, *Post-Task* twice (once for each of the tasks). The SPIRE system automatically balanced the task order. No data was acquired in the *Introduction*, *Pre-Task*, and *Thank you* steps.

- Select one book about surviving on a desert island
- Select one book that will teach you something new
- Select one book about one of your personal hobbies or interests
- Select one book that is highly recommended by other users (based on user ratings and reviews)
- Select one book for fun

Please add a note (in the book-bag) explaining why you selected each of the five books.

The *Non-goal* Task. was developed based on the open-ended task used in the iCHiC task at CLEF 2013 [10] and the ISBS task at CLEF 2014 [4]. 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 following instruction text was provided to participants:

> Imagine you are waiting to meet a friend in a coffee shop or pub or the airport or your office. While waiting, you come across this website and explore it looking for any book that you find interesting, or engaging or relevant. Explore anything you wish until you are completely and utterly bored. When you find something interesting, add it to the book-bag. Please add a note (in the book-bag) explaining why you selected each of the books.

5.2 Experiment Structure

The experiment was conducted using the SPIRE system[6] [5], using the flow shown in Figure 2. Each participant ran through the *Pre-Task*, *Task*, *Post-Task* steps once for each of the two tasks. When a new participant started the experiment, the SPIRE system automatically allocated them to one of the two tested interfaces and to a given task order. Interface allocation and task order were automatically balanced to minimise bias in the resulting data. 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 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 [8] engagement scale.

5.3 System and Interfaces

The two tested interfaces (*baseline* and *multi-stage*) were both built using the PyIRE[7] 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.

Both interfaces used a shared IR backend implemented using ElasticSearch[8], which provided free-text search, faceted search, and access to the individual books complete metadata. 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

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

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

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

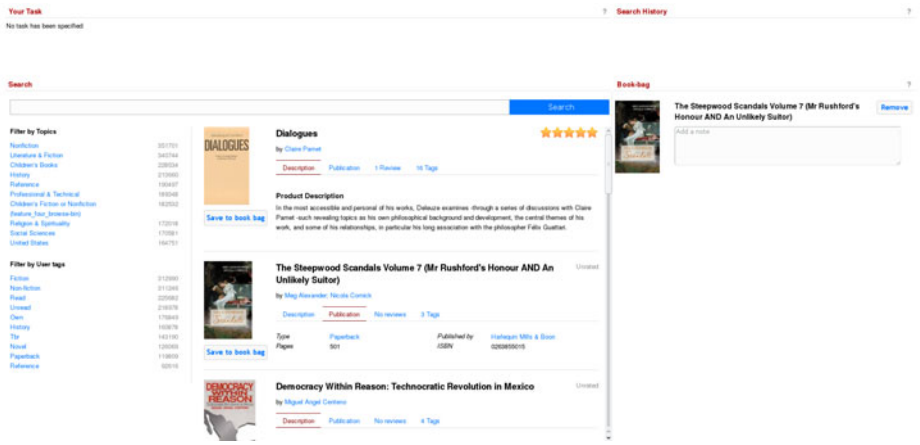

Fig. 3. *Baseline* interface – uses a standard faceted search interface, consisting of a search box, search facets based on the Amazon subject classifications and user tags, and the book-bag on the right.

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 *Baseline* Interface. shown in figure 3 represents a standard faceted web-search interface, the only additions being the task information (top-left) and the list of past searches (top-right). The main interface consists of a search box at the top, two facets on the left, and the search results list (center). On the right-hand side is the book-bag, which shows the participants which books they have collected for their task and also provides the notes field, which the participants were instructed to use to explain why they had chosen that book.

The two facets provided on the left use the Amazon subject classification and the user tags to generate the two lists together with numeric indicators for how many books each facet contained. Selecting a facet restricted the search results to books with that facet. Participants could select multiple facets from both lists. In the search results list each book consisted of a thumbnail image, title, authors, aggregate user rating, a description, publication information (type, publisher, pages, year, ISBN ...), user reviews, and user tags (where available). The aggregate user rating was displayed using 1 to 5 stars in half-star steps, calculated by aggregating the 1-5 star ratings for each user review. If the book had no user reviews, then no stars were shown. Additionally each book had a "Add to Bookbag" button that participants used to add that book into their bookbag.

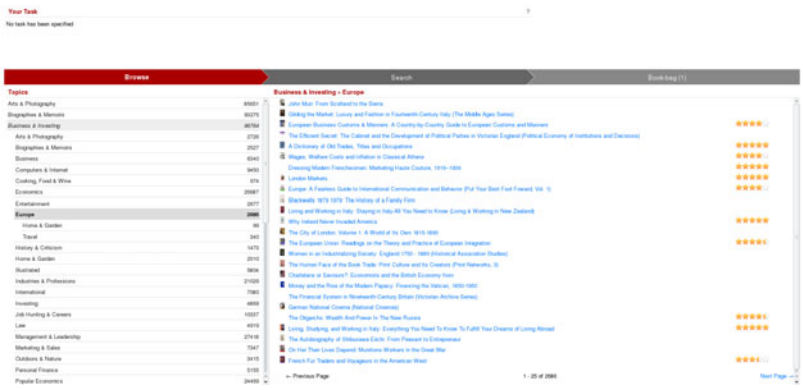

Fig. 4. *Multistage* interface – *Browse* view – subject browse hierarchy derived from the Amazon subject classifications on the left and the dense search results list on the right with thumbnail, title, and aggregate ratings for each book.

The *Multi-stage* Interface. aims to support users by taking the different stages of the search process into account. The idea behind the *multi-stage* interface design is supported by two theoretical components.

Firstly, several information search process models look at stages in the search process. A well-known example is [7], who discovered "common patterns in users' experience" during task performance. She developed a model consisting of six stages, which describe users' evolving thoughts, feelings and actions in the context of complex tasks. [11] later summarized Kuhlthau's stages into three categories (pre-focus, focus formulation, and post-focus), and points to the types of information searched for in the different stages. The multi-stage search interface constructed for iSBS was inspired by [11]. 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.

Secondly, when designing a new search interface for social book search it has also been relevant to look more specifically at the process of choosing a book to read. A model of decision stages in book selection [9] identifies the following decision stages: browse category, selecting, judging, sampling, and sustained reading. This work supports the need for a user interface that takes the different search and decision stages into account. However, the different stages in [9] closely relate to a specific full text digital library, and therefore the model was not applicable to the present collection.

When the *multi-stage* interface first loads, participants are shown the *browse* stage (fig. 4), which is aimed at supporting the initial exploration of the dataset. 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 [3], which uses the relative frequencies of the subjects to arrange them into the tree-structure

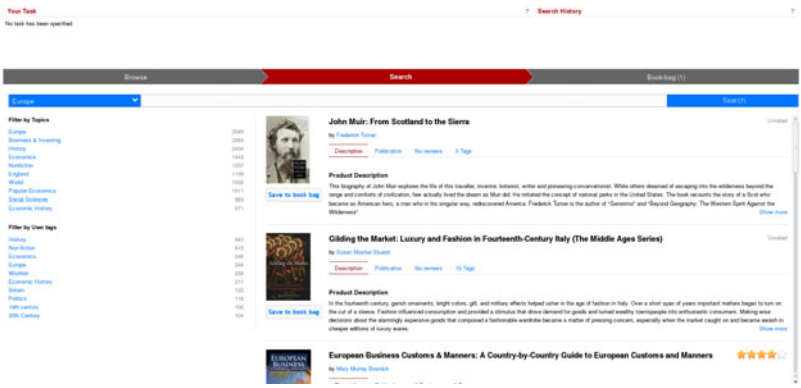

Fig. 5. *Multistage* interface – *Search* view – faceted search interface that matches the interface used in the *Baseline* interface. Differences are the inclusion of the Amazon subject selection box next to the search box and the removal of the book-bag.

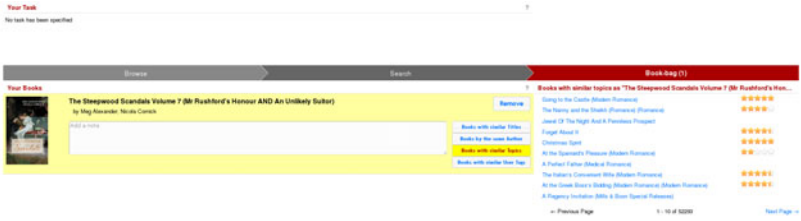

Fig. 6. *Multistage* interface – *Book-bag* view – books added to the book-bag are listed on the left together with the note areas for each book. On the right the list of similar books using the dense result list from the *Browse* view.

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 using the same layout and content as used in the *baseline* interface's search result list.

Participants switched to the *search* stage by clicking on the "Search" section in the gray bar at the top. The *search* stage (fig. 5 uses the same interface as the *baseline* with only two differences. The first is that as the book-bag is a separate stage, it is not shown on the *search* stage interface itself. The second is that if the participants select a topic in the *browse* stage, this topic is pre-selected as a filter for any queries in the blow box to the left of the search box. Participants can click on that box to see a drop-down menu of the selected topic and its parent topics. Participants can select a higher-level topic to widen their search.

The final stage is the *book-bag* shown in Figure 6, where participants review the books they have collected and can provide notes for each book. For each

book, buttons were provided that allow the user to search for similar books by title, author, topic, and user tags. The similar books are shown on the right using the same compact layout as in the *browse* stage. As in the *browse* stage, clicking on a book in that list shows a popup window with the book's details.

5.4 Participants

A total of 192 participants were recruited (see Table 1), 120 female and 72 male. 72 were between 18 and 25, 80 between 26 and 35, 25 between 36 and 45, 8 between 46 and 55, 6 between 56 and 65 and 1 over 65. 60 were in employment, 3 unemployed, 128 were students and 1 selected *other*. Participants came from 36 different countries (country of birth) including Germany (63 participants), UK (33), Denmark (21), Norway (20), the Netherlands (11), resident in 13 different countries, again mainly in Germany, UK, Denmark, Norway and the Netherlands. Participants mother tongues included German, Dutch, English, Danish, Romanian, Farsi, Portuguese and 23 others. The majority of participants executed the tasks remotely (136) and only 56 users in a lab. 95 participants used the novel *multi-stage* interface, while 97 used the *baseline* interface.

5.5 Procedure

Participants were invited by the individual teams, either using e-mail or by recruiting students from a lecture or lab. Where participants were invited by e-mail, the e-mail contained a link to the online experiment, which would open in the participant's browser. Where participants were recruited in a lecture or lab, the experiment URL was distributed using e-learning platforms. 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, which mostly focused on gathering additional feedback on the system from the participants.

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 automatically using JavaScript and stored with the participants' responses. Table 3 shows median and inter-quartile ranges for all interface and task combinations. Session lengths are significantly lower for the *baseline* interface (wilcoxon signed rank $p < 0.05$). Also all session lengths are significantly longer than in the iSBS 2014 experiment [4].

Table 3. Session lengths, number of queries executed, and number of books collected for the two interfaces and tasks. Times are in minutes:seconds, numbers reported are median and inter-quartile range.

Interface	Goal-oriented		Non-goal	
	Median	inter-quartile	Median	inter-quartile
Session length				
Baseline	10:30min	10:25min	5:33min	7:37min
Multi-Stage	12:52min	9:20min	7:18min	10:52min
Number of queries				
Baseline	8	5	2	3
Multi-Stage	6	6.5	1	3
Number of books				
Baseline	5	0	3	3
Multi-Stage	5	0	3	3

Number of Queries. was extracted from the log-data. In both interfaces it was possible to issue queries 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 interface and task. The number of queries per session is significantly higher for the *baseline* interface over the *multi-stage* interface for both tasks (wilcox $p < 0.05$) and also for the *goal-oriented* over the *non-goal* task in both interfaces (wilcox $p < 0.01$).

Number of Books Collected. was extracted from the log-data. Participants collected those books that they felt were of use to them. The numbers reported in Table 3 are based on the number of books participants had in their book-bag when they completed the session, not the total number of books collected over the course of their session, as participants could always remove books from their book-bag in the course of the session.

Unlike the other metrics, there is no significant difference between the two interfaces. On the *goal-oriented* task this was expected as participants were asked to collect five books. On the *non-goal* task this indicates that the interface had no impact on what participants felt were enough books to complete the task.

6 Conclusions and Plans

This was the first year of the SBS Lab, which is a continuation from the SBS and iSBS Tracks at INEX 2014. The overall goal remains to investigate the relative value of professional metadata, user-generated content and user profiles. The number of active participants increased in both tracks, from 8 to 11 in the Suggestion Track and from 4 to 7 for the Interactive Track, indicating there is strong interest in the IR community for research in the domain of books and social media.

In the Suggestion Track, the setup was mostly the same as in 2014. Topic statements have both a natural language narrative of the information need and one or more books provided as positive or negative examples of what the user is looking for. In addition to the explicitly marked up book suggestions in the forum threads, We included manually extracted suggestions that were not marked up. With the examples participants can investigate the value of query-by-example techniques in combination with more traditional text-based queries. 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. Next year, we continue this focus on complex topics with example books and consider including an recommendation-type evaluation. We also consider extending the task by asking systems 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.

The interactive track investigated how searchers make use of and appreciate professional metadata and user-generated content for book search on the Web. Two interfaces were tested to identify and analyse the different stages in the search process. This was the second year of the Interactive Track, in which we improved the two interfaces to identify and analyse the different stages in the search process in the domain of book search. One interface resembles traditional search interfaces familiar from Amazon and LibraryThing, the other is a *multi-stage* interface where the first part provides a broad overview of the collection, the second part allows the user to look at search results in a more detailed view and the final part allows the user to directly compare selected books in great detail. This year seven teams collaborated to get a shared data pool of 192 participants from many different backgrounds and countries. We found that users spent significantly more time searching the multistage interface than the baseline interface but issued fewer queries, probably because the multistage interface allows browsing as an extra mode of exploring the collection. For the next year, we plan to have multiple experiments focused on specific research questions, with fewer users per experiment. Another option is to let individual teams plan their own experiments.

One possibility for synergy between the two tracks that we intend to investigate next year is how to define experiment tasks that will enable the comparison of results and approaches between the two tracks. Sharing tasks would allow us to evaluate results from the *Suggestion* track based on the users' performances in the *Interactive* track. Another possibility could be to investigate whether some of the successful (re-)ranking techniques used in the Suggestion track could be implemented in the search engine used in the Interactive track.

References

1. Beckers, T., Fuhr, N., Pharo, N., Nordlie, R., Fachry, K.N.: Overview and results of the INEX 2009 interactive track. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 409–412. Springer, Heidelberg (2010)
2. Borlund, P., Ingwersen, P.: The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation **53**(3), 225–250 (1997)
3. Hall, M.M., Fernando, S., Clough, P.D., Soroa, A., Agirre, E., Stevenson, M.: Evaluating hierarchical organisation structures for exploring digital libraries. Information Retrieval **17**(4), 351–379 (2014)
4. Hall, M.M., Huurdeman, H.C., Koolen, M., Skov, M., Walsh, D.: Overview of the INEX 2014 interactive social book search track. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) Working Notes for CLEF 2014 Conference. CEUR Workshop Proceedings, Sheffield, UK, September 15–18, 2014, vol. 1180, pp. 480–493. CEUR-WS.org (2014)
5. Hall, M.M., Toms, E.: Building a common framework for IIR evaluation. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 17–28. Springer, Heidelberg (2013)
6. Koolen, M., Kamps, J., Kazai, G.: 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)
7. Kuhlthau, C.C.: Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science **42**(5), 361–371 (1991)
8. O'Brien, H.L., Toms, E.G.: 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)
9. Reuter, K.: Assessing aesthetic relevance: Children's book selection in a digital library. JASIST **58**(12), 1745–1763 (2007)
10. Toms, E., Hall, M.M.: The chic interactive task (chici) at clef2013 (2013). http://www.clef-initiative.eu/documents/71612/1713e643-27c3-4d76-9a6f-926cdb1db0f4
11. Vakkari, P.: 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)