INEX 2012 Social Book Search Track
Search Task Definitions and Submission Guidelines

Marijn Koolen, Gabriella Kazai and Michael Preminger Version 3.0

Introduction

The goal of the Social Book Search Track is to evaluate techniques that support users in searching, navigating and reading collections of book descriptions and digitized books. The track investigates two tasks: Social Book Search (SBS), Prove It! (PI).

Participation

To take part you will need to register to the book track at
https://inex.mmci.uni-saarland.de/people/register.jsp.

Once you registered, you will be issued with a username and password. These will give you access to the data (book corpus, topics, relevance judgements) and services on http://www.booksearch.org.uk/. If you encounter any problems, please email Gabriella Kazai at v-gabkaz@microsoft.com.

Participants may take part in either of the tasks. The minimum participation requirement is to contribute results to the SBS or the PI task.

The deadline for submitting runs is 15 June for the Social Book Search task and 22 June for the Prove It task.

What's new in 2012 compared to 2011?

- This year, the Social Book Search (SBS) will focus on specific aspects of the relevance of books and descriptions. We will not only compare the LibraryThing forum suggestions with the relevance assessments from Amazon Mechanical Turk (AMT), but also look at which forum suggestions are subsequently catalogued by LibraryThing (LT) members who created the topics. In the AMT assessment phase, we will ask assessors to look at topical relevance, reading level, fun-factor or interestingness and recommendation.
- In addition, the topics will not only contain the topic statements from the discussion threads, but also the personal profile and catalogue from the topic creator. Participants may exploit this user context information to improve search results. This adds a recommendation aspect to the search task.
- For the PI task the topics from last year will be given more structure. The complex factual claims will be split into a list of atomic parts for evaluation, such that assessors can indicate which part of a claim is confirmed or refuted by a book page. Evaluation will take into account which parts of a claim are supported or discredited by information in books. Also, recall will be a more important aspect of evaluation, which means a larger part of submitted runs will be used to create assessment pools.
- Submission format has been changed to TREC style format, with a 7th column for the confirm/refute label.

Social Book Search (SBS)

1.1 SBS goals

The aim of the SBS task is to investigate book search within the context of professional and user-generated data on books. Book search in library catalogues is traditionally restricted to formal and subject access points, allowing users to search on title information or subject terms from controlled vocabularies. Users of Amazon and LibraryThing provide additional information about books in the form of ratings, reviews and tags, which often goes beyond the subject of a book into describing writing style, reading level, comprehensiveness and engagement. This allows not only to extend the searchable data to user-generated content, but also allows users to search for a broader range of information needs. The goal is to evaluate the relative value of *controlled book metadata* versus *user-generated content* for retrieving the most relevant books for a broad range of user requests.

Controlled metadata, such as the Library of Congress Classification and Subject Headings, is rigorously curated by experts in librarianship. It is used to index books to allow highly accurate retrieval from a large catalogue. However, it requires training and expertise to use effectively, both for indexing and for searching. On the other hand, user-generated content is less rigorously defined and applied, and lacks vocabulary control. However, such metadata is contributed directly by the users and it may better reflect the terminology of everyday searchers. Clearly, both types of metadata have advantages and disadvantages. The task aims to investigate whether one is more suitable than the other to support different types of search requests and how they may be fruitfully combined.

The task addresses the following research questions.

Request-related questions:

- What types of book requests do users of the LibraryThing discussion forums have?
- How are book requests related to the objectives and functions of library catalogues?
- How are book requests related to professional metadata (traditional catalogue access points)?
- How are book requests related to user-generated content?

Task-related questions:

- How are book suggested on the LibraryThing discussion forums related to the book request?
- What criteria do searchers use to select or ignore suggested books?

System-related questions:

- What is the relative value of professional metadata and user-generated content for social book search?

1.2 SBS user scenario

The scenario is that of a user turning to Amazon Books and LibraryThing to search for books they want to gather on a given topic. Both services host large collaborative book catalogues that may be used to locate books of interest.

On LibraryThing, users can catalogue the books they read, manually index them by assigning tags, and write reviews for others to read. Users can also post messages on a discussion forum asking for help in finding new, fun, interesting, or relevant books to read. The forums allow users to tap into the collective bibliographic knowledge of hundreds of thousands of book enthusiasts.

On Amazon, users can read and write book reviews and browse to similar books based on links such as ‛customers who bought this book also bought… ‛.

1.3 SBS task description

The SBS task is to reply to a user's request that has been posted on the LibraryThing forums by returning a list of recommended books. The books must be selected from a corpus that consists of a collection of book metadata extracted from Amazon Books and LibraryThing, with additional metadata from the British Library and the Library of Congress. The collection includes both curated and social metadata. User requests vary from asking for books on a particular genre, looking for books on a particular topic or period or books by a given author. The level of detail also varies, from a brief statement to detailed descriptions of what the user is looking for. Some requests include examples of the kinds of books that are sought by the user, asking for similar books, or list examples of known books that are related to the topic but are specifically of no interest. Other requests describe the reading level, writing style or type of story of the books they're looking for. The challenge is to develop a retrieval method that can cope with such diverse requests.

Participants may submit up to 6 runs and can use any field of the topic statement. However, **at least one run should use the title field only**.

1.4 SBS corpus

The corpus consists of a collection of 2.8 million records from Amazon Books and LibraryThing.com. See https://inex.mmci.uni-saarland.de/data/nd-agreements.jsp for information on how to get access to this collection. Each book record is an XML file with fields like <isbn>, <title>, <author>, <publisher>, <dimensions>, <numberofpage> 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 comes from Amazon and LibraryThing, in the <tag>, <rating>, and <review> fields.

There are two additional sets of official library records for a subset of the ISBNs from Amazon and LibraryThing. These library records come from the British Library and the Library of Congress, and contain official library classification information such as Library of Congress Classifications (LCC) and Subject Headings (LCSH). Participants are not required to use these library records. They are merely provided as additional professional metadata, as not all Amazon/LibraryThing records have classification codes and subject headings, and some of the Amazon subject headings are inappropriate. The library records increase the number of books with professional metadata and possibly improve the quality as well.

1.5 SBS topics

Last year we released a small set of training topics for the SBS task,. For this year, training material consists of the same set of 43 training topics and the 211 official topics of last year. Both are available on the INEX website. For more detail on the topic sets and relevance judgements, see the INEX 2011 overview paper of the Books and Social Search Track.

This year's official topic set consists of 301 topics taken from the LibraryThing discussion forums, and includes the 211 topics from last year. However, this year, we focus more on the actual user catalogues of the topic creators and take into account which of the suggested books in the topic threads are added by the topic creator to her personal catalogue. We believe this is a stronger signal of relevance than considering all suggested books as equally relevant. We have crawled the user catalogues of the topic creators, which contains both the dates on which books were added to the catalogue and the tags that the user assigned to them.

- The books that the topic creator added **before** she started the topic thread will be made available as context information for the search topics. Participants may use this context information to improve retrieval and ranking. This provides a recommendation aspect to the task.
- The books added **after** the thread was started will be filtered on the suggestions in the thread. The suggestions from the thread added after starting the topic will be considered the most relevant suggestions.

Topics consist of the following fields:

- title = the title of the topic
- user = username of the topic creator (to be used with the user profiles that are related separately)
- group = the group name in which the topic is discussed
- narrative = the first message of the topic, posted by the topic creator
- type = the topic type, manually added by organisers. Type can be subject, author, genre, series or known-item
- genre = the genre of the topic (fiction, non-fiction or both).

As an example topic, here is the format for topic 99309:

```
<topic id="99309">
  <title>Politics of Multiculturalism</title>
  <user>steve.clason</user>
  <group>Political Philosophy</group>
    <narrative>I'm new, and would appreciate any recommended reading on the politics of multiculturalism. <author>Parekh</author>'s <work id="164382">Rethinking Multiculturalism: Cultural Diversity and Political Theory</work>(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 <author>Sen</author>, <author>Rawls</author>, <author>Habermas</author>, and <author>Nussbaum</author>, 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>
  <type>subject</type>
  <genre>non-fiction</genre>
</topic>
```

The topic and work id's are taken directly from LibraryThing. The mapping from the LibraryThing work id to the ISBNs in the collection is made using thingISBN.xml, which is compiled by LibraryThing and can be obtained from:

http://www.librarything.com/feeds/thingISBN.xml.gz

The user profiles have the following fields:

- username name of the user
- about_library short description of the user's library
- friends list of usernames of friends on libraryThing.
- groups list of groups that the user is a member of
- interesting_libraries list of usernames that the user considers to have interesting libraries
- library list of books that the user has added to her catalogue, including date of entry and ratings and tags
- reviews reviews of books that the user has written

The topics and profiles can be downloaded from:

https://inex.mmci.uni-saarland.de/data/documentcollection.jsp#books

1.6 SBS submission format

Submissions should conform to the TREC format of:

<topic-id> Q0 <ISBN> <rank> <retrieval-score> <run-id>

A run should contain up to 1,000 books per topic. Entries should be sorted by topic-id and then by retrieval-score. We will use ISBN numbers as book identifiers – these are provided as part of the test collection. Run IDs must be unique across all submissions sent from one organization – also please use meaningful, but short names if possible. Please note that we will use trec_eval at the evaluation stage to calculate performance scores, which ignores the rank column and orders documents by retrieval score.

1.7 SBS evaluation

Systems will be evaluated using relevance judgements extracted from the LibraryThing topic threads and LibraryThing user catalogues of the topic creators. Books suggested in the topic threads will be treated as relevant, and the number of forum members suggesting the same book will be used as the basis for the relevance grade. Suggested books that the topic creator adds to her library after creating the topic are considered the most relevant books. Additional evaluation will be based on crowdsourced judgements from Amazon Mechanical Turk. The overall performance across topic types will be measured using nDCG@10. Performance for different topic types will be individually evaluated using specific measures. For known-item topics we will use MRR and Success@10, for subject-related topics we will use nDCG@10 and MAP.

Prove It! (PI)

1.8 PI goals

Building on a corpus of over 50,000 digitized books, this task investigates the application of focused and semantic retrieval approaches to a collection of digitized books.

1.9 PI user scenario

The scenario underlying this task is that of a user searching for specific information in a library of books that can provide evidence to *confirm or refute a given factual statement*. Users expect to be pointed directly at book pages that can help them to confirm or refute the claim of the topic. Users are assumed to view the ranked list of retrieved book pages starting from the top of the list and moving down, examining each result. No browsing is considered (only the returned book pages are viewed by users).

1.10 PI task description

Systems need to return a ranked list of up 1,000 book pages (given by their XPaths) which are estimated to contain information that can confirm or refute a factual statement expressed in the topic, and (optionally) indicating for each result whether it provides positive or negative evidence regarding the claim.

Participants may submit up to 6 runs.

1.11 PI corpus

The PI task builds on a collection of 50,239 out-of-copyright books, digitized by Microsoft. The corpus contains books of different genre, including history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry. 50,099 of the books also come with an associated MAchine-Readable Cataloging (MARC) record, which contains publication (author, title, etc.) and classification information. Each book in the corpus is identified by a 16 character bookID, which is also the name of the directory that contains the book's OCR file, e.g., A1CD363253B0F403.

The OCR text of the books has been converted from the original DjVu format to an XML format referred to as BookML, developed by Microsoft Development Center Serbia. BookML provides additional structure information, including a set of labels (as attributes) and additional marker elements for more complex structures, like table of contents. For example, the first label attribute in the XML extract below signals the start of a new chapter on page 1 (label=" PT_CHAPTER"). Other semantic units include headers (SEC_HEADER), footers (SEC_FOOTER), back-of-book index (SEC_INDEX), table of contents (SEC_TOC). Marker elements provide detailed markup, e.g., indicating entry titles (TOC_TITLE) or page numbers (TOC_CH_PN) in a table of contents. The basic XML structure of a typical book in BookML is a sequence of pages containing nested structures of regions, sections, lines, and words, most of them with associated coordinate information, defining the position of a bounding rectangle.

The full corpus, totalling around 400GB, is available on USB HDDs – please contact Gabriella Kazai at v-gabkaz@microsoft.com. A reduced version (13GB compressed) is available via download from http://www.booksearch.org.uk/. The reduced version was generated by removing the word tags and propagating their content into the parent (i.e., line) elements.

1.12 PI topics

1.13

There are 83 topics for this task, each containing a factual statement, see example below.

```
<topic id=2010006 ct_nt=11>

  <fact>On November 2, 1917, British prime minister Lord Balfour, in a letter to Lord Rotschild, declared the British Government's intention of assisting the Jewish people to establish a national home for the jews in Palestine.</fact>
  <query>Balfour declaration national home</query>
  <subject>national home for jews in Palestine</subject>
  <wikiurl>http://en.wikipedia.org/wiki/Balfour_Declaration_of_1917</wikiurl>
  <narrative>
    <task> I am working on a school assignment about the establishment of the state of Israel as the fulfillment of the Zionist aspirations. I was told that the Balfour declaration was an important milestone, and I want the fact confirmed. </task>
    <infneed> All statements in books are relevant that either prove or reject the fact. I need confirmation of the exact date, and preferably also the text of the declaration. </infneed>
  </narrative>
</topic>
```

Topics are available from the download area at http://www.booksearch.org.uk/.

1.14 PI training data

We have page-level relevance judgements for 21 of the 83 topics, which can be used for training. The judgements can be downloaded from the download area at http://www.booksearch.org.uk/.

1.15 PI evaluation

This is a high-precision task, with graded judgements, thus the official measure will be nDCG@10. Most topics contain complex factual statements consisting of multiple atomic facts that each can be confirmed or refuted on a page. During the assessment phase, pages will be judged on each of these atomic facts, which makes judging easier and the interpretation of the judgements clearer. For a topic with n atomic facts, a page can thus confirm or refute up to n of these facts. The number of confirmed/refuted facts on a page determine the relevance level. That is, a page confirming/refuting *n* facts has a relevance level of *n*.

1.16 Submission format

Submissions should conform to an extended version of the TREC format:

```
<topic-id> Q0 <bookID-pageno> <rank> <retrieval-score> <run-id> <label>
```

The 3rd column is a combination of the bookID and the page number. **Please note that page numbers are not the same as the value of the id attribute in the <page id="..."> elements. The first page element is page number one. This coincides with the numbering of the page elements according to xpath.**
The 7th column is an extra column which should indicate whether the result at that rank confirms or refutes the factual statement of the topic. Possible values are confirm, refute or both.

```
2010000 Q0 4BCB760E4F1A4E4D-59 0 -3.29573 0 Confirm
```

As an example, the result line above states that for topic 2010000, page 59 of the book with ID 4BCB760E4F1A4E4D has a retrieval score of -3.29573 and confirms the statement according to the retrieval system,

How to submit your runs

Runs should be uploaded to http://booksearch.org.uk. Login using your INEX credentials and go to the Upload Area. Please note that no online validation of your runs is offered, so please make sure that your runs conform to the submission format described above.