RSLIS at INEX 2013: Social Book Search Track

Toine Bogers and Birger Larsen

Royal School of Library and Information Science
University of Copenhagen
Birketinget 6, 2300 Copenhagen, Denmark
{tb,blar}@iva.dk

Abstract. In this paper, we describe our participation in the INEX 2013 Social Book Search track. We investigate the contribution of different types of document metadata, both social and controlled, as well as the contribution of the new 'query' topic representations. We find that the best results are obtained using all available document fields and topic representations.

Keywords: XML retrieval, social tagging, controlled metadata, book recommendation

1 Introduction

In this paper, we describe our participation in the INEX 2013 Social Book Search track[1]. Our goal for the Social Book Search task was to investigate the contribution of the new 'query' topic field representation provided for this year's task.

The structure of this paper is as follows. We start in Section 2 by describing our methodology: pre-processing the data, which document and topic fields we used for retrieval, and our evaluation. In Section 3, we describe the results of our content-based retrieval runs, including the effect of the additional topic field representation. Section 4 describes which runs we submitted to INEX, with the results of those runs presented in Section 5. We discuss our results and conclude in Section 6.

2 Methodology

2.1 Data and Preprocessing

In our experiments we used the Amazon/LibraryThing collection provided by the organizers of the INEX 2012 Social Book Search track. This collection contains XML representations of 2.8 million books, with the book representation data crawled from both Amazon.com and LibraryThing (LT). The 2012 collection is identical to the collection provided for the 2011 track [1] in all but two ways: the collection has been expanded with additional library records from the British Library (BL) and the Library of Congress (LoC). Of the 2.8 million books in the collection, 1.15 million

[1] https://inex.mmci.uni-saarland.de/tracks/books/

have a BL record and 1.25 have a LoC record. Together these two sources cover 1.82 million of the 2.8 million books in the collection.

We converted the collection's original XML schema into a simplified version to retain only those metadata fields that were most likely to contribute to the successful retrieval of relevant books[2]. After these pre-processing steps, we were left with the following 19 content-bearing XML fields in our collection: <isbn>, <title>, <publisher>, <editorial>, <creator>, <series>, <award>, <character>, <place>, <blurber>, <epigraph>, <firstwords>, <lastwords>, <quotation>, <dewey>, <subject>, <browseNode>, <review>, and <tag>.

We replaced the numeric Dewey codes in the original <dewey> fields by their proper textual descriptions using the 2003 list of Dewey category descriptions[3] to enrich the controlled metadata assigned to each book. For example, the XML element <dewey>519</dewey> was replaced by the element <dewey>Probabilities & applied mathematics</dewey>. The BL and LoC records were provided in MODS format[4], we mapped this format to the appropriate new XML fields and added them to the book representations.

2.2 Field categories and Indexing

The 19 selected XML fields in our collection's book representations fall into different categories. Some fields, such as <dewey> and <subject>, are examples of *controlled metadata* produced by LIS professionals, whereas other fields contains *user-generated metadata*, such as <review> and <tag>. Yet other fields contain 'regular' book metadata, such as <title> and <publisher>. Fields such as <quotation> and <firstwords> represent a book's content more directly.

To examine the influence of these different types of fields, we divided the document fields into five different categories, each corresponding to an index. To examine the contribution of the additional BL/LoC controlled metadata we created two versions of the index containing controlled metadata: one with and one without this additional controlled metadata. In addition, we combined all five groups of relevant fields for an index containing all fields. This all-fields index also comes in two variants: one with and one without the BL/LoC metadata. This resulted in a total of eight indexes:

All fields For our first index all-doc-fields we simply indexed all of the available XML fields (see the previous section for a complete list). The all-doc-fields-plus index contains all of the original 2011 fields as well as the BL/LoC metadata.

Metadata In our metadata index, we include all metadata fields that are immutably tied to the book itself and supplied by the publisher: <title>, <publisher>, <editorial>, <creator>, <series>, <award>, <character>, and <place>.

Content For lack of access to the actual full-text books, we grouped together all XML fields in the content index that contain some part of the book text: blurbs,

[2] Please consult [2] for more details on this filtering and conversion process.

[3] Available at http://www.library.illinois.edu/ugl/about/dewey.html

[4] See http://www.loc.gov/standards/mods/ for more information.

epigraphs, the first and last words, and quotations. This corresponded to indexing the fields <blurber>, <epigraph>, <firstwords>, <lastwords>, and <quotation>.

Controlled metadata In our controlled-metadata index, we include the three controlled metadata fields curated by library professionals harvested from Amazon: <browseNode>, <dewey>, and <subject>. The controlled-metadata-plus index contains the original metadata as well as the BL/LoC metadata.

Tags We split the social metadata contained in the document collection into two different types: tags and reviews. For the tags index, we used the tag field, expanding the tag count listed in the original XML. For example, the original XML element <tag count="3">fantasy</tag> would be expanded as <tag>fantasy fantasy fantasy</tag>. This ensures that the most popular tags have a bigger influence on the final query-document matching.

Reviews All user reviews belonging to a single book were combined in a single document representation for that book and added to our review index reviews.

We used the Indri 5.1 retrieval toolkit[5] for indexing and retrieval. We performed stopword filtering on all of our indexes using the SMART stopword list, and preliminary experiments showed that using the Krovetz stemmer resulted in the best performance. Topic representations were processed in the same manner.

2.3 Topics

As part of the INEX 2013 Social Book Search track three sets of topics were released with requests for book recommendations based on textual description of the user's information need: three training sets and a test set. All topic sets were extracted from the LibraryThing forum. The original training set of 43 topics created for the 2011 Social Book Search track came with unverified relevance judgments. The 2011 Social Book Search track also supplied participants with 211 topics with relevance judgments derived from the books recommended on the LibraryThing discussion threads of these 211 topics. We used these 2011 test set as our training set to optimize our retrieval algorithms in the different runs. The results we report in Sections 3 and **??** were obtained using this training set.

The test set for 2013 contains 386 new topics which were used to rank and compare the different participants' systems at INEX 2013. The results listed in Section 5 were obtained on this combined set of 386 topics. Each topic is represented by several different fields:

Title The <title> field contains the title of the forum topic and typically provide a concise description of the information need. Runs that only use the topic title are referred to as title.

Group The LibraryThing forum is divided into different groups covering different topics.

[5] Available at http://www.lemurproject.org/

Narrative The first message of each forum topic, typically posted by the topic creator, describes the information need in more detail. This often contains a description of the information need, some background information, and possibly a list of books the topic creator has already read or is not looking for. The narrative typically contains the richest description of the topic.

Query In 2013, three annotators hired to classify aspects of the 386 topics were also asked to provide query representations for each topic.

All topic fields We also performed runs with the title, group and narrative fields combined, referred to as all-topic-fields.

All topic fields + query We also performed runs with all four individual fields combined, referred to as all-plus-query to test the contribution of the query field.

In our experiments with the training and the test set, we restricted ourselves to automatic runs using the query, title, all-topic-fields, and all-plus-query representations (based on our experiments for INEX 2011-2012 [2]).

2.4 Experimental setup

In all our retrieval experiments, we used the language modeling approach with Jelinek-Mercer (JM) smoothing as implemented in the Indri 5.1 toolkit. We preferred JM smoothing over Dirichlet smoothing, because previous work has shown that for longer, more verbose queries JM smoothing outperforms Dirichlet smoothing [3], which matches the richer topic descriptions provided in the topic sets.

For the best possible performance, we optimized the λ parameter, which controls the influence of the collection language model, with higher values giving more influence to the collection language model. We varied λ in steps of 0.1, from 0.0 to 1.0 using the training set of topics. We also examined the value of stop word filtering and stemming and use the SMART stop word list and Krovetz stemming in these cases. This resulted in 44 different possible combinations of these three parameters. For each topic we retrieved up to 1000 documents and we used NDCG@10 as our evaluation metric [4].

3 Content-based Retrieval

In order to produce a competitive baseline for our experiments, we conducted a first round of experiments focused on optimizing a standard content-based retrieval approach for all index and topic representations[6]. We found that the best results were always produced with stop word filtering and Krovetz stemming, so all results reported in this paper share these settings. We compared the different index and the different topic representations for a total of 16 different content-based retrieval runs. Table 1 shows the best NDCG@10 results for each run on the training set.

[6] This did not include the query and all-query-plus representations as the 2011 test set did not include the query topic field.

Table 1. Results of the 16 different content-based retrieval runs on the training set using NDCG@10 as evaluation metric. Best-performing runs for each topic representation are printed in bold.

Document fields	Topic fields	
	title	all-topic-fields
metadata	0.0915	0.2015
content	0.0108	0.0115
controlled-metadata	0.0406	0.0496
controlled-metadata-plus	0.0514	0.0691
tags	0.0792	0.2056
reviews	0.1041	0.2832
all-doc-fields	**0.1129**	**0.3058**
all-doc-fields-plus	0.1120	0.3029

We can see several interesting results in Table 1. First, we see that the best overall content-based run used all topic fields for the training topics, retrieved against the index containing all document fields (all-doc-fields) with an NDCG@10 score of 0.3058. Retrieving on the all-doc-fields index performs best on both topic sets (all-topic-fields and title), so for the runs submitted to the 2013 Social Book Search track we restricted ourselves to the all-doc-fields index.

4 Submitted runs

We selected three automatic runs for submission to INEX[7] based on the results of our content-based runs.

Run 1 (query.all-doc-fields) This run used the query field of the test topics and ran this against the index containing all available document fields (with the exception of the additional BL/LoC metadata).

Run 2 (all-topic-fields.all-doc-fields) This run used the title, narrative, and group fields of the test topics and ran this against the index containing all available document fields (with the exception of the additional BL/LoC metadata).

Run 3 (all-plus-query.all-doc-fields) This run used all four topic fields and ran this against the index containing all available document fields (with the exception of the additional BL/LoC metadata).

5 Results

The runs submitted to the INEX 2013 Social Book Search track were evaluated using graded relevance judgments[8]. All runs were evaluated using NDCG@10, P@10,

[7] Our participant ID was 54.

[8] The rules for assigning the different relevance grades can be found at https://inex.mmci.uni-saarland.de/tracks/books/results.jsp.

MRR, with NDCG@10 as the main metric. Table 2 shows the official evaluation results.

Table 2. Results of the three submitted runs on the test set, evaluated using all 386 topics with relevance judgments extracted from the LibraryThing forum topics. The best run scores are printed in bold.

Run #	Run description	NDCG@10	P@10	MRR
1	query.all-doc-fields	0.0401	0.0208	0.0635
2	all-topic-fields.all-doc-fields	0.1295	0.0647	0.2190
3	all-plus-query.all-doc-fields	**0.1361**	**0.0653**	**0.2286**

We see that, unsurprisingly, the best-performing run on all 386 topics was run 3 with an NCDG@10 of 0.1361. Run 3 used all available topic fields including the new query field and all document fields.

6 Discussion & Conclusions

On both the training and the test sets the best results were achieved by combining all topic and document fields. This shows continued support for the principle of polyrepresentation [5] which states that combining cognitively and structurally different representations of the information needs and documents will increase the likelihood of finding relevant documents.

References

1. Kazai, G., Koolen, M., Kamps, J., Doucet, A., Landoni, M.: Overview of the INEX 2011 Book and Social Search Track. In: INEX 2011 Workshop pre-proceedings. INEX Working Notes Series (2011) 11–36
2. Bogers, T., Christensen, K.W., Larsen, B.: RSLIS at INEX 2011: Social Book Search Track. In Geva, S., Kamps, J., Schenkel, R., eds.: INEX 2011: Proceedings of the 10th International Workshop of the Initiative for the Evaluation of XML Retrieval. Volume 7424 of Lecture Notes in Computer Science., Berlin, Heidelberg, Springer Verlag (2012) 45–56
3. Zhai, C., Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Information Retrieval. ACM Transactions on Information Systems **22**(2) (2004) 179–214
4. Järvelin, K., Kekäläinen, J.: Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems **20**(4) (2002) 422–446
5. Ingwersen, P.: Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. Journal of Documentation **52**(1) (1996) 3–50