SOCIAL BOOK SEARCH TRACK: ISM@INEX'14 SUGGESTION TASK

Ritesh Kumar and Sukomal Pal

Department of Computer Science and Engineering,
Indian School of Mines Dhanbad, 826004
India
{ritesh4rmrvs,sukomalpal}@gmail.com

Abstract. This paper describes the work that we did at Indian School of Mines towards Social Book Search Track for INEX 2014. We submitted five runs in its Suggestion Task. We investigated individual effect of *title*, *group*, *mediated_query*, and *narrative* fields of the topics in our runs. For all the runs we used language modelling technique with Dirichlet smoothing. The run using only *mediated_query* field was our best. Overall, our performance is not satisfactory. However, as new entrant to the field, our scores are encouraging enough to work for better results in future.

Keywords: Book Search, Social Book Search, Language modelling, Information Retrieval

1 Introduction

With growing numbers of online portals and book catalogues, our current time sees a rapid evolution in the way we acquire, share and use books. In order to enable users, Social Book Seach Track at INEX [5] provides a relevant experimental platform to investigate techniques of searching and navigating professional metadata provided by publishers/booksellers and user-generated content from social media [1]. At INEX 2014, they offered two tasks: Suggestion Task and Interactive Task. We participated in the first where we were supposed to recommend books based on user's request and her personal catalogue data (list of books with rating and tags maintained for the user in the social cataloguing site). We were also provided with a large set of anonymised user profiles from LibraryThing forum members. Each user request is provided in the form of topics containing different fields like *title*, *mediated_query*, *group*, *narrative* and catalogue information.

As a newcomer to this field, our goal this year was to investigate the contribution of different topic fields in book recommendation. We only considered *title*, *mediated_query*, *group*, *narrative* fields from each topic. We did not consider topic-creator's catalogue information. Neither we consulted anonymous user profiles.

We submitted five runs (run-ids: ISMD-341, ISMD-342, ISMD-350, ISMD-354, ISMD-355) in the Suggestion Task. For all the runs, Language modelling

with Dirchlet smoothing was used in Lemur's Indri search system [3]. Our overall performance was not satisfactory. The run with only *mediated_query* was best among our submissions.

Organization of rest of the paper is as follows. We describe our approach in Section 2. Section 3 describes dataset and Section 4 reports results. In Section 5 we analyse our results. Finally, we conclude in Section 5 with directions for future work.

2 Approach

This year we took a simple approach similar to standard adhoc retrieval. The document collection provided was stopword-removed and then stemmed using Krovetz stemmer. It was indexed with Lemur Indri search system for all the fields having text within.

During retrieval, we tried to see the effect of different components of a topic in turn. We therefore used only *title* (Run-id ISMD-341), only *group*(Run ISMD-342), only *title with stopword removed* (Run ISMD-350), only *mediated_query* (Run ISMD-354), and only *narrative with stopword removed* field (Run ISMD-355) from each topic.

On top of standard English stopwords we identified a set of a few more like *recommendation*, *hello*, *suggestion*, *reference*, *recent*, *hi*, *thank*, etc. which we removed in the run ISMD-355.

We also removed punctuation marks manually from all the textual content of these fields and used only free text queries in all the runs.

We did not consider any other information like catalogue information and user profile during retrieval.

For each topic, we submitted 1000 book suggestions in the form of ISBNs.

3 Data

Test collection provided by INEX 2014 SBS orgainzers for Suggestion Task had a document collection and a topicset. The document collection consists of 2.8 million book description with metadata from Amazon and LibraryThing. From Amazon there is formal metadata like booktitle, author, publisher, publication year, library classification codes, Amazon categories and similar product information, as well as user-generated content in the form of user ratings and reviews. From LibraryThing, there are user tags and user-provided metadata on awards, book characters and locations and blurbs. There are additional records from the British Library and the Library of Congress. The entire collection was 7.1 GB in size. [2]

The topic-set contains 681 topics each describing a user's request for suggestion of books. Each topic has a set of fields like *title*, *mediated_query*, *group*, *narrative* and user's personal catalogue at the time of topic creation. The catalogue contains a list of book-entries with information like LibraryThing id of the book, its entry-date, rating and tags.

The organizers also supplied 94,000 anonymised user profiles from LibraryThing.

4 Results

The scores obtained by our five runs are given in Table 1. The official evaluation measure by INEX'14 is nDCG@10 [4]. The performance of our runs are in decreasing order. Our best performance is by ISMD-354 where we use only *mediated_query* field. We also show the best score in the task demonstrated by run-id **USTB-run6.SimQuery1000.rerank_all.L2R_RandomForest**(*), for the sake of comparison.

Table 1. Results - The official evaluation Measure by INEX 2014

RUN ID	Rank	MRR	**nDCG@10**	MAP	R@1000
ISMD-354	22	0.123	**0.067**	0.049	0.285
ISMD-341	24	0.106	0.056	0.042	0.236
ISMD-350	27	0.090	0.048	0.036	0.211
ISMD-355	29	0.089	0.038	0.026	0.124
ISMD-342	32	0.018	0.010	0.007	0.081
*best**	1	0.464	0.303	0.232	0.390

5 Analysis

Although our performance is not up to the mark, there are few take-home lessons. As individual fields, *mediated_query* is the most effective, followed by *title* and *narrative*. Removing stopwords from the *title* is actually detrimental (ISMD-341 and ISMD-350). We did not consider any combination of these fields. It would be interesting to see the performance of different combinations of these fields.

6 Conclusion

This year we participated in the Suggestion Task of Social Book Search as initial venture. We tried to see the individual effect of different topic-fields on book recommendation. We considered only a handful of fields like *mediated_query*, *title*, *narrative* etc from the topics. While there can be no denial of the fact that our overall performance is dismal, initial results are suggestive as to what should be done next. We need to consult other fields like book catalogue of the topic creators, ratings of the books in the catalogue during retrieval. We also need to take into account profiles of other users. It is also imperative to see the performance of combination of different fields in the topics as well as other fields in user catalogues and user profiles. We shall be exploring some of these tasks in the coming days.

References

1. Marijn Koolen, Gabriella Kazai, Jaap Kamps, Michael Preminger, Antoine Doucet and Monica Landoni, *Overview of the INEX 2012 Social Book Search Track*. INEX'12 Workshop Pre-proceedings, Shlomo Geva, Jaap Kamps, Ralf Schenkel (editors), September 17-20, 2012, Rome , Italy.
2. INEX, Initiative for the Evaluation of XML Retrieval. https://inex.mmci.uni-saarland.de/data/documentcollection.jsp
3. INDRI: Language modeling meets inference networks, Available at http://www.lemurproject.org/indri/
4. Jarvelin, K., Kekalainen, J.: Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems 20(4) (2002) 422-446.
5. INEX, Initiative for the Evaluation of XML Retrieval. https://inex.mmci.uni-saarland.de/