A Methodology for Social Book Search

VenkataRaviKiran Ravva, Lakshmi Lavanya Singampalli, Vamshi Krishna Thotempudi, Carolyn J. Crouch

Department of Computer Science
University of Minnesota Duluth
Duluth, MN 55812
(218) 726-7607
ccrouch@d.umn.edu

Abstract. A general overview of our methodology and results for the INEX 2014 Social Book Search Suggestion Task are presented in this paper. This is our first entry in the Social Book Search Track, which started in 2011. Our methodology and experiments are inspired by background research on the Social Book Search Track [5, 6, 7, 8, and 9]. We originally submitted six runs to the INEX 2014 competition and subsequently expanded our experiments as time allowed. Results, though preliminary, indicate some positive directions for future examination.

1 Introduction

Books have always been prominent sources of information. The Social Book Search Track [2] was introduced by INEX in 2011 with the purpose of providing support to users in terms of easy search and access to books using metadata. In 2014 the track includes two tasks: the suggestion task and the interactive task. We worked on the suggestion task, which suggests a ranked list of books to satisfy the user's query. The goal is to compare results generated via a traditional system which uses professional and social metadata to that of a recommender system which uses user preference information.

Much of this year's focus is on the recommender system. With the aim of retrieving more relevant books for a query, we designed a recommender system that uses similar users as its basis for grouping books. The nearest neighbors are used along with the cosine similarity measure to generate a set of similar users associated with each topic. An aspect of particular interest here is the use of both user-generated and professional metadata. The method combines the two aspects of retrieval and recommendation.

Indexing of documents is done using the search engine Indri [11], and evaluation measures are calculated using prescribed TREC metrics. The scores obtained from traditional retrieval are further manipulated and re-ranked to produce a new set of recommended scores. These results are again submitted to TREC evaluation and any improvements in the evaluations are recorded. In this paper we provide a brief

summary of the methodology followed, the experiments conducted and the results obtained. This paper describes both our traditional and recommender systems in their current state.

2 Overview

The data made available by INEX includes the following. The topics are divided into six different groups based on various combinations of the title, query, group and narrative tags. The six topic groups are title, query, title-query, title-query-group, title-query-narrative and title-query-group-narrative. The traditional metadata (both professional and social) from the Amazon corpus and LibraryThing are indexed using Indri to retrieve a ranked set of documents for each query. Six different parses (social data, Amazon data, full [all tags], LibraryThing, professional data, and title) are generated to produce six different indices. Results are subjected to TREC evaluation to produce final results.

Our methodology includes a recommender system that uses user profile content to find users similar to the user who originally posted the query. Once the similar users are found, new, recommender scores are generated by using a linear combination of the traditional score and the recommender-generated score for that query. Based on these scores, the original, ranked list of documents is re-ranked to produce the final set of results. The linear combination used to generate the final scores is based on the parameter λ, which is critical in the re-ranking process. The top 1000 books for each query are retrieved and returned as results.

3 Experiments

Returning a ranked set of relevant documents in response to a user's query is the goal of the Social Book Search (SBS) Track [2]. The input data set is comprised of 2.8 million documents that contain information from both Amazon [4] and LibraryThing (LT) [3]. Each document is in XML format with social, professional and user-generated metadata. The 680 queries include not only content (i.e., a statement of user need) but also information about the user's catalog. Apart from these two data sets, approximately 94,000 anonymous user profiles are provided to each participating team for experimental purposes (for the recommender system experiments, in particular).

Our approach combines two methods, namely, retrieval and recommendation. Both are described below.

3.1 Traditional System

The traditional system is responsible for document retrieval, including scrubbing, parsing, and indexing using Indri. Six different indices are generated, based on differ-

ent contents from the input XML. These are the social, professional, LT, full, Amazon, and title indices. The queries are also processed based on their XML tags. Each query has four XML tags: title, mediated-query, group, and narrative. We use the following query sets for our experiments: Title (T), Query (Q), Title-Query (TQ), Title-Query-Group (TQG), Title-Query-Narrative (TQN) and Title-Query-Group-Narrative (TQGN).

We began our experiments by testing our approach and methodology on the 2013 data [1] which was available to us. Using each index and query set, retrieval was performed both with and without pseudo-relevance feedback [10] to produce an initial ranked list of documents. Specified numbers of documents and of terms were used to perform pseudo feedback using Indri, with the number of documents, **d**, ranging from 5 to 15, and the number of terms, **t**, ranging from 15 to 50. Best results were produced using the full index (as indicated by [8] for the 2013 data). Using feedback values of **d**=10 and **t**=50 with the TQG query set produced the highest nDCG@10 value. We selected this run as the basis for our 2014 submitted runs. That is, the run produced by pseudo-feedback (**d**=10, **t**=50) on the **full** index with the **TQG** query set is used as our basic retrieval run for INEX submission.

Upon access to the 2014 QRels, we re-examined our feedback values of **d** and **t** with the aim of improving recall. R@1000 improves from 0.328 (at **d**=10 and **t**=50) to 0.380 at **d**=10 and **t**=15, as seen in Table 1. We use this retrieval run as the basis for our current results.

Table 1. Results of Traditional Retrieval (Full Index, TQG Query Set with Pseudo-feedback)

Run	# docs	#terms	nDCG@10	MRR	MAP	R@1000
Official INEX run	10	50	0.095	0.185	0.068	0.328
Current results	10	15	0.091	0.182	0.064	0.380

3.2 Recommender System

This is the second stage of the system, where the results produced by traditional retrieval are re-ranked by the recommender system. The recommender system is designed to make use of information from users "similar to" the user who posted the query. Here we assume that similar users tend to have similar preferences and tastes in books.

3.2.1 Finding Similar Users

The first step in our recommender system generates a matrix for each of the 680 topic users. These matrices consist of work IDs and tags, because we want to examine and

identify similar books (work IDs) and similar genres (tags). The values in the matrices are combinations of numeric and binary values. We consider that users must have a minimum of 5 work IDs in common before they are considered similar. The matrix representations are presented in Table 2.

Table 2. Matrix Representation

Matrix Representation	Work ID Value	Tag Value
bin_bin	binary 1 = work ID exists 0 = otherwise	binary 1 = tag exists 0 = otherwise
bin_num	binary 1 = work ID exists 0 = otherwise	numeric tag frequency
num_bin	numeric rating for work ID	binary 1 = tag exists 0 = otherwise
num_num	numeric rating for work ID	numeric rating for work ID

Once the matrices are generated, the next step is to generate a list of similar users based on the context vectors. Pairwise cosine similarity is used as the similarity measure. The top-ranked 50 and 100 "similar users" are considered the sets of interest.

3.2.2 Generating the Contribution of the Recommender System

We now generate Δ, the contribution of recommender system, using as input, for each primary user: (1) the rank ordered list of similar users, (2) the similarity score of each such user, (3) the rating for each work ID identified by document retrieval, and (4) the count of similar users having that same work ID in their catalogs. Here we use 2 metrics to calculate Δ, the contribution of the recommender system. One metric employs a DCG-style metric, and the other uses an MRR approach. These metrics are defined in Table 3.

3.2.3 Generating Final Scores

A linear combination of the scores produced by traditional retrieval and Δ, the contribution of the recommender system, produces a re-ranked list of "recommended" documents. By fine tuning λ, we arrive at a value of 0.0000075 for Metric 1 and 0.0000125 for Metric 2. The results obtained by both evaluations are presented in Table 4.

Table 3. Metrics for Calculating Δ (the Contribution of the Recommender System)

Metric	Binary Score	Numeric Score
Metric 1 (DCG-style)	$R_{ij} = \sum_{k=1}^{50/100} \frac{S_{ik} + 1}{\log_2^{rank} + 1}$	$R_{ij} = \sum_{k=1}^{50/100} \frac{S_{ik} + r_{jk}}{\log_2^{rank} + 1}$
Metric 2 (MRR- style)	$R_{ij} = \sum_{k=1}^{50/100} \frac{S_{ik} + 1}{rank}$	$R_{ij} = \sum_{k=1}^{50/100} \frac{S_{ik} + r_{jk}}{rank}$
i = topic id j = work ID k = similar user for topic 'i' (50/100) R_{ij} = Recommended score for topic 'i' work ID 'j' S_{ik} = Similarity score for user 'k' r_{jk} =Rating given by user 'k' for work ID 'j'		

Table 4. Final Results of the Recommender System

Metric	Feature	Users	λ	nDCG@10	MRR	MAP	R@1000
Metric 1	bin_num	50	0.0000075	0.0965	0.1931	0.0662	0.3801
		100	0.0000075	0.0958	0.1932	0.0661	0.3801
	bin_bin	50	0.0000075	0.1025	0.2041	0.0715	0.3801
		100	0.0000075	0.1004	0.1997	0.0697	0.3801
Metric 2	bin_num	50	0.0000125	0.0977	0.1946	0.0670	0.3801
		100	0.0000125	0.0978	0.1961	0.0685	0.3801
	bin_bin	50	0.0000125	0.1058	0.2077	0.0746	0.3801
		100	0.0000125	0.1053	0.2084	0.0722	0.3801

4 Analysis and Conclusions

The best features for the matrices are binary_binary (bin_bin), where both work IDs and tags are represented as binary values. The value of nDCG@10 is greater when 50 rather than 100 similar users are considered. Metric 2 produces a higher nDCG@10 result.

We note here that many relevant documents are not being retrieved by traditional retrieval. Increasing recall at this stage may be expected to produce improvement in the final scores. Our current best result (0.1058) would rank at 17 in terms of nDCG@10 and 13 in terms of R@1000 when compared to the INEX 14 official results. Many of these results exhibit small differences; we do not yet know if they are significant. This is our first attempt at this task, which has proved to be an excellent learning experience.

References:

1. About INEX [Internet]. Amsterdam, Netherlands. INEX: c 2008-2013 [cited 2013 Jun 20]. Available from: https://inex.mmci.uni-saarland.de/about.html
2. About INEX 2014 Social Book Search Track. Available from: https://inex.mmci.uni-saarland.de/tracks/books/
3. About LibraryThing [Internet]. Available from: https://www.librarything.com/about
4. Amazon website: http://www.amazon.com/
5. Adriaan, F., Kamps, J. and Koolen, M.: University of Amsterdam at INEX 2011: Book and Data Centric Tracks, *INEX 2011 Workshop Pre-proceedings*, INEX Working Notes Series, pp.36-48, 2011.
6. Bogers, T. and Larsen, B.: RSLIS at INEX 2012: Social Book Search Track. *INEX 2012 Workshop Pre-proceedings*, INEX Working Notes Series, pp. 36-48, 2012.
7. Bogers, T., Wilfred, C. K. and Larsen, B.: RSLIS at INEX 2011: Social Book Search Track. In: S. Geva, J. Kamps, and R. Schenkel (Eds.): *INEX 2011*, LNCS 7424 , pp. 45-56, 2012.
8. Huurdeman, H., Kamps, J., Koolen, M. and Wees, J. V.: Using Collaborative Filtering in Social Book Search, *INEX 2012 Workshop Pre-proceedings*, INEX Working Notes Series, pp.125-136, 2012.
9. Koolen, M., Kazai, G., Preminger, M. and Doucet, A.: Overview of the INEX 2013 Social Book Search Track, *CLEF 2013 Working Notes Series*, Valencia, Spain, 2013.
10. Lavrenko, J. and Croft, W. B.: A language modeling approach to information retrieval, *Proceedings of the 1998 ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–281, 1998.
11. Strohman, T., Metzler, D., Turtle, H. and Croft, W. B.: Indri: A language model-based search engine for complex queries, *Proceedings of the 2nd International Conference on Intelligence Analysis*: 2(6), pp. 2-6, 2005.