Comparing Topic Representations for Social Book Search

Marijn Koolen[1], Hugo Huurdeman[1,2], and Jaap Kamps[1,2,3]

[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[3] ISLA, Faculty of Science, University of Amsterdam

Abstract. In this paper we describe our participation in the INEX 2013 Social Book Search Track. We compare the impact of different query representations for book search topics derived from the LibraryThing discussion forums, including the title and full narrative provided by the topic creator, the name of the discussion group in which the topic was posted, and a mediated search query provided by a trained annotator. Our findings are that 1) the mediated queries are short and do not improve performance over the titles, but combining titles and mediated queries does, 2) the discussion group name adds relevant new terms to the representation and further improves performance, but adding the narrative is not effective, and 3) for the majority of topics retrieval effectiveness is the same across all topic representations. Our findings suggest that writing a good search query for the complex information needs in social book search is far from trivial, even for trained annotators.

1 Introduction

For the INEX 2013 Social Book Search Track we focused our attention on query representations. The search topics in this track are based on discussion threads from the LibraryThing (LT) discussion forums and contain both the title of the topic threads, the narrative in the first message of the thread and a mediated query created by a trained annotator. The latter one is provided by the track organisers to compensate for non-representative thread titles for some of the forum topics.

The topic statements of the SBS Track contain rich representations of the book search information needs. The LT member who starts the topic thread describes her information need both in the thread title and in detail in the first message of the thread. In addition, she choses a discussion group in which to start the thread, which broadly categorises her information need, with the aim to attract responses from LT members who are knowledgeable about relevant books and can recommend the best ones.

These different representations may each reflect different aspects of the information need. In our participation we investigate how these representations affect retrieval. Specifically, we want to know:

– How different are the thread title and the mediated query and how does that affect retrieval performance?
– What is the importance of the detailed narrative, that explains the information need in detail, for representing the information need?
– What is the role of the discussion group name in representing the information need?

In addition, we experiment with a document prior based on the book ratings of LT members. We crawled a large set of user profiles from LT that includes which book each member added to her catalogue and the rating she assigned to it. The average rating of a book may reflect its overall quality, in which case it could be used to push low quality and non-rated (and therefore unpopular) books down the ranking.

The paper is structured as follows. We first discuss the different topic representations that are available in this year's topic set in Section 2. Then, we describe our experimental setup in Section 3 and discuss results in Section 4. Next, in Section 5 we present a per-topic analysis. In Section 6, we discuss our findings and draw conclusions.

2 Topic Representations

The topics for the SBS task are based on topic threads on the LT discussion forums. Each thread starts with a message from the topic creator and is posted in one of the thousands of discussion groups. The 2013 topic set only contains topic threads that are started with a book search information need. The thread has a title and the first message can be seen as a narrative of the information need. For instance, topic 25244 has title *Why Republic vs. Democracy* and is posted in the *Political Conservatives* discussion group. The narrative explains that the user wants to know more about forms of government and the logic behind choosing one or the other.

What is a good topic representation to use as a search query? The title is often a concise summary of the information need, but is not always comprehensive, especially for very detailed needs. Sometimes titles are conversational and reveal nothing about the topic of the information need, such as topic 45940 with title *Request for recommendations....* The narrative explains the user is looking for books about the miracles of Jesus that are not based on the Bible. The title is a bad representation of the information need, while the narrative contains much more than just the information need. Because these titles and narrative are not intended as search engine queries, this year the task organisers provided a mediated search query with each topic, created by a trained annotator. This query is meant to be both concise and comprehensive.

We want to investigate the value of this mediated query with respect to the thread title and the narrative. Does it provide a better representation than the thread title? Does it cover all the fine-grained aspects expressed in the narrative? And what is the role of the Group name of the discussion group that the user

Table 1: Statistics on the number of words per topic representation for different combinations on topic fields (T = thread title, Q = mediated query, G = group name and N = narrative).

		Total words					Distinct words				
Fields	# tpcs	min.	max.	med.	mean	std.dev	min.	max.	med.	mean	std.dev
T	386	0	12	4	3.90	2.07	0	12	4	3.88	2.04
Q	386	1	10	3	3.61	1.55	1	9	3	3.56	1.47
TQ	386	2	19	7	7.51	3.00	1	15	5	5.77	2.46
TQG	386	4	20	9	9.46	2.86	2	16	7	7.19	2.39
TQN	386	4	257	43	53.68	39.20	3	179	32	39.93	27.96
$TQGN$	386	6	259	44	55.63	39.04	4	179	32	41.03	27.69

selected? This group broadly categorises the information need with, we assume, the aim to find LT members who are knowledgeable on books about the subject. But it may also be useful as an additional representation of the topic.

The topic set contains 386 topics and each topic has five fields: *title* (T), *query* (Q), *group* (G), *member* and *narrative* (N). We ignore the member field, which contains the name of the topic creator and is probably not useful for representing the information need. To understand some of the differences between these fields as possible topic representations, we analyse them in terms of the number query terms they contain.

In Table 1 we see statistics on the number of query terms in (combinations of) fields, based on the text in those fields after parsing, stopword removal and Krovetz stemming. This processing corresponds to the way documents are processed before indexing. Columns 3–7 show the total number of content terms and columns 8–12 show the number of distinct terms. The title field (T) has a mean (median) of 3.90 (4) content terms. The number of distinct terms is very similar, showing that content terms are rarely repeated in the title. There is one topic, number 28304, which has zero content terms, for which the thread title is *Who am I? Why am I here?*. This is a topic posted in the *Amateur Historians* group asking about books on exploration. Apart from the title containing only highly frequent words, it also does not reflect the information need at all. Here the mediated query, *exploration books*, improves the query representation. The query field (Q) is in general somewhat shorter—the median is 3—but there is always at least one content term. This poses the question whether the mediated query is more comprehensive than the title, reflecting aspect from the narrative not covered in the title. Again, terms in the field are rarely repeated. The combination of the T and Q fields results in an almost doubling of the number of content terms. The number of distinct query terms is lower but still higher than the number of distinct terms in either the title or query field. This means that many but not all of content terms in the title and query overlap. It is plausible that the most relevant terms from the title are repeated in the query, which results in higher term frequencies for the most important terms. This might be beneficial for retrieval.

Next, we add the group and narrative fields to the combined title and query field. The group adds only one or two terms on average, while the narrative adds dozens of content terms, with some repeated terms. However, the narrative usually contains some conversational language, with many content terms not directly related to the information need. It is not clear to what extent the possibly larger number of relevant content terms can increase performance and to what extent its conversational distractor terms hurt performance.

In Section 4 we discuss how these different fields affect retrieval effectiveness.

3 Experimental Setup

We used Indri [4] for indexing, removed stopwords and stemmed terms using the Krovetz stemmer. Based on the results from the 2011 Social Search for Best Books task [1] we include all the social metadata. From the Amazon/LibraryThing (A/LT) collection we use the booktitle, author name, subject headings, LT tags and Amazon user reviews for indexing. In addition, we use the Library of Congress Subject Headings (LCSH) from the catalogue records of the British Library and the Library of Congress. These subject headings are less noisy than the headings from Amazon, and there are more headings per book.

The topics are taken from the LibraryThing discussion groups and contain a *title* field which contains the title of a topic thread, a *group* field which contains the discussion group name and a *narrative* field which contains the first message from the topic thread. New this year is a mediated *query* field, which is provided by the organisers as an additional representation of the information need and is meant to be a more precise expression of it than the thread title.

In our experiments we used different combinations of topic fields as queries. For the language model our baseline has default settings for Indri (Dirichlet smoothing with $\mu = 2500$). We created six base runs:

T : a standard LM run using only the Title field of the topic.
Q : a standard LM run using only the Query field of the topic.
TQ : a standard LM run using the Title and Query fields of the topic.
TQG : a standard LM run using the Title, Query and Group fields of the topic.
TQN : a standard LM run using the Title, Query and Narrative fields of the topic.
TQGN : a standard LM run using the Title, Query, Group and Narrative fields of the topic.

Last year we crawled a large set of user profiles from LT members and used member catalogues and book ratings to rerank retrieval results based on nearest-neighbourhood recommendation. This year, we use the Bayesian average book ratings as document priors. That is, books that received ratings from LT members are boosted up the ranking with respect to books that received no ratings and books with high ratings are boosted more than books with low ratings.

To normalise the ratings, we compute the Bayesian average of all the book ratings in the top 1000 results per topic. The Bayesian Average (BA) takes into

account how many users have rated a work. As more users rates the same work, the average becomes more reliable and less sensitive to outliers. We make the BA dependent on the query, such that the BA of a book is based on books related to the query. The BA of a book b is computed as:

$$BA(b) = \frac{\hat{n} \cdot \hat{m} + \sum_{r \in R(b)} r}{n + \hat{n}} \tag{1}$$

where $R(b)$ is the set of ratings for b $\hat{m}$ is the average unweighted rating over all books in the top 1000 results and $\hat{n}$ is the average number of ratings over all the books in the top 1000.

A rating $BA(b)$ for book b can range from 0.5 up to 5, with increments of 0.5. For books with no rating we use $BA = 0$. a base score of 1, for books with ratings we use $1 + BA$. Each rating can be turned into a prior probability by dividing BA by the maximum rating $BA_{max} = 5$. For books with no rating this would results in a prior probability of zero. To avoid multiplying by zero, we use the Add-One smoothing method and compute the prior as:

$$P_{BA}(d) = \frac{1 + BA(d)}{1 + BA_{max}} \tag{2}$$

The final document score is then:

$$S_{BA}(d) = P(d|q) \cdot P_{BA}(d) \tag{3}$$

We submitted six runs:

inex13SBS.ti_qu : the TQ run.
inex13SBS.ti_qu_gr_na : the TQGN run.
inex13SBS.ti.bayes_avg.LT_rating : the T run with the Bayes LT rating prior.
inex13SBS.qu.bayes_avg.LT_rating : the Q run with the Bayes LT rating prior.
inex13SBS.ti_qu.bayes_avg.LT_rating : the TQ run with the Bayes LT rating prior.
inex13SBS.ti_qu_gr_na.bayes_avg.LT_rating : the TQGN run with the Bayes LT rating prior.

In the next section we discuss the evaluation results of the official submission and separately all our own runs.

4 Results

We first show the evaluation results over the whole topic set. Then we present a per-topic analysis of the differences in performance between the different topic representations.

Table 2: Evaluation results of the top 10 runs of the INEX 2013 SBS task. Our runs are in italics

Rank	Run ID	nDCG@10	P@10	MRR	MAP
1	run3.all-plus-query.all-doc-fields	0.1361	0.0653	0.2286	0.0861
2	*inex13SBS.ti_qu_gr_na.bayes_avg.LT_rating*	0.1331	0.0771	0.2342	0.0788
2	*inex13SBS.ti_qu.bayes_avg.LT_rating*	0.1331	0.0771	0.2342	0.0788
4	run1.all-topic-fields.all-doc-fields	0.1295	0.0647	0.2190	0.0797
5	*inex13SBS.ti_qu_gr_na*	0.1184	0.0555	0.2075	0.0790
6	*inex13SBS.ti_qu*	0.1163	0.0647	0.2091	0.0665
7	run_ss_bsqstw_stop_words_free_member_free_2013	0.1150	0.0479	0.1839	0.0800
8	run_ss_bsqstw_stop_words_free_2013	0.1147	0.0468	0.1843	0.0798
8	*inex13SBS.qu.bayes_avg.LT_rating*	0.1147	0.0661	0.1997	0.0656
10	*inex13SBS.ti.bayes_avg.LT_rating*	0.1095	0.0634	0.2005	0.0630

Table 3: Evaluation results of our runs in the INEX 2013 SBS task. Significance levels are 0.05 ($^\circ$), 0.01 ($^\circledcirc$) and 0.001 ($^\bullet$).

Run ID	nDCG@10	%	P@10	%	MRR	%	MAP	%
T	0.094		0.053		0.190		0.066	
Q	0.097	2.6%	0.054	1.3%	0.187	-1.3%	0.065	-1.7%
TQ	0.116$^\circledcirc$	23.3%	0.065$^\bullet$	21.6%	0.220$^\circledcirc$	15.9%	0.082$^\circledcirc$	24.2%
TQG	**0.120**$^\bullet$	**27.4**%	**0.068**$^\bullet$	**27.1**%	**0.225**$^\circledcirc$	**18.7**%	0.084$^\circledcirc$	28.3%
TQN	0.115°	21.6%	0.052	-3.0%	0.204	7.9%	0.089$^\circledcirc$	35.8%
$TQGN$	0.118°	25.6%	0.056	4.3%	0.217	14.5%	**0.093**$^\circledcirc$	**40.9**%
TP_{BA}	0.110		0.063		0.209		0.077	
Q_{BA}	0.115	4.8%	0.066	4.3%	0.209	0.1%	0.079	2.9%
TQ_{BA}	0.133$^\bullet$	21.6%	**0.077**$^\bullet$	**21.6**%	0.244$^\circledcirc$	16.7%	0.095$^\bullet$	23.4%
TQG_{BA}	**0.135**$^\bullet$	**23.4**%	**0.077**$^\bullet$	**22.1**%	**0.247**$^\circledcirc$	**18.3**%	0.096$^\bullet$	24.1%
TQN_{BA}	0.132°	20.5%	0.063	-0.3%	0.237	13.6%	**0.102**$^\circledcirc$	**31.9**%
$TQGN_{BA}$	0.132°	20.6%	0.067	5.4%	0.235	12.7%	0.100$^\circledcirc$	29.1%

This year, eight groups participated in the track submitting a total of 32 runs. Our official submissions are all among the top 10 systems, as shown in Table 2. The top four systems are close together in terms of performance, as are the systems on ranks five up to nine. Our systems perform on par with the best other systems.

We show the evaluation results of our own runs in Table 3. Significant differences are tested using the bootstrap method (one-tailed with 100,000 samples). Significance levels are 0.05 ($^\circ$), 0.01 ($^\circledcirc$) and 0.001 ($^\bullet$). In the top half of the table we see the base runs without Bayes Average ratings priors. Significance tests are with respect to the title-only (T) run. Somewhat surprisingly, the title-only (T) and query-only (Q) representations lead to similar performance. The mediated query does not improve the representation of the information need. However, the combination of title and mediated query (TQ) gives significantly better performance than either in isolation. This reflects the fact that the query is not

Table 4: Per topic differences in nDCG@10 between runs

	# topics		
	↓	=	↑
$S(Q) - S(T)$	74	237	69
$S(TQ) - S(T)$	50	256	74
$S(TQ) - S(Q)$	49	255	76
$S(TQG) - S(TQ)$	53	257	70
$S(TQN) - S(TQ)$	84	222	74
$S(TQGN) - S(TQ)$	81	220	79

simply a copy of the thread title, but either adds complementary relevant terms or gives more weight to the most relevant terms by repeating them, or both.

Adding the group name to the title and query (TQG) further improves performance, reflecting the users ability to pick relevant discussion groups for their needs. However, adding the more detailed narrative hurts performance for early precision (nDCG@10, P@10 and Mean Reciprocal Rank (MRR)) while improving Mean Average Precision (MAP). It seems the narrative is not focused enough to precisely pinpoint the suggested books but its larger set of query terms does lead to better recall.

In the bottom half of Table 3 we the six runs with Bayes Average rating priors. Again, significant differences are with respect to the title-only T_{BA}. The rating priors lead to improvements on all reported measures for all six baseline runs. Among the runs with rating priors we see the same patterns as among the baseline runs. The T and Q representations lead to similar performance but their combination leads to better performance. The group name improves the topic representation but the narrative hurts early precision while improving MAP. We also tested the improvements of the prior ratings runs over their baseline forms and found that all improvements are significant for $p < 0.001$, except for the TQGN run where the improvements are significant for $p < 0.05$. This shows the reliability of the rating priors.

In sum, the title and query representations are equally effective but complementary to each other. The group name can further improve performance while the narrative seems to add too many partly relevant and irrelevant terms. The LT ratings, if normalised by taking the Bayesian average, forms a reliable document prior probability of relevance.

5 Per-Topic Analysis

We show the per topic differences between two runs for nDCG@10 in Table 4. The Q run has lower scores for 74 topics compared to the T run (column 2), higher scores for 69 topics and the same scores as the T run for 237 topics. These two runs are balanced, which explains why they lead to similar average scores, but the large number of topics for which the two runs get the same score suggests that in most cases the mediated query is very similar to the thread

title. It also suggests that creating an effective representation of the information need is far from trivial, even for trained annotators. Some of their mediated queries improve upon thread titles that do not or only partly reflect the often complex information needs in social book search [3]. But even more mediated queries express the search topic less well than the title created by the topic creator. Next we compare the TQ with the T and Q runs (rows 3 and 4). These are less balanced, with TQ outperforming T on 74 topics and Q on 76 topics while T outperforms TQ for only 50 topics and Q outperforms TQ for 49 topics. This explains why the combination of the two representations scores higher on average than either on its own. Because T and Q are often very similar, their combination also often results in the same score.

Finally, we compare the per topic scores of the TQ representations with the richer representations TQG, TQN and TQGN. The TQG run improves performance on more topics than on which it decreases performance, which corresponds with an improvement on the average score. The representations that include the narrative, TQN and TQGN, both worsen performance with respect to the TQ representations on more topics than on which they improve, corresponding to a drop in performance in nDCG@10. What is surprising is that including the much longer narrative in the representation does not affect the per topic score for the majority of topics. There are several possible explanations for this. It could be that additional terms often provide the same relevance signal as the TQ terms, or introduce a random noise. Another explanations is that the TQ terms are frequently repeated in the narrative and therefore have a dominant impact on the retrieval score.

To summarise, the different query representations often carry the same signal, which may be because the same content terms dominate in the representations. However, it seems hard to improve upon the title created by the topic starter, but combining the concise representations of topic starter and annotator more often results in an improved representation than in a worse one.

6 Conclusion

In this paper we discussed our participation in the INEX 2013 Social Book Search Track in which we focus on the impact of different query representations of the information needs on retrieval effectiveness. The LT members who start a topic thread to ask for book suggestions on the discussion forums provide multiple types of perspectives on their information needs. The thread title is a short summary, the first message in the thread is a detailed description and the choice of the particular discussion group reveals the relevant general category of books for which they hope to find knowledgeable members. In addition the task organisers provided mediated queries that aim to be both concise and comprehensive expressions of the information need, and that are suitable as search engine queries.

The mediated query in general slightly shorter than the thread title, and typically contains a few overlapping terms and one or a few different content

terms. By combining the representations, the overlapping terms in the title and query—which we assume are the most relevant terms—receive extra weight.

The group name is short but also tends to add a few new terms to the representation with respect to the title and query. The narrative is much longer and adds many terms, relevant or not to the representation.

In terms of the impact of representations on retrieval effectiveness, the title and mediated query are equally effective. Their combination, however, leads to significant improvements over using the title alone, which is either due to the higher frequency of the most important terms or to the complementary content terms. However, for most topics, the title, query and their combination lead to the same retrieval performance. Adding the group name improves performance, indicating that the user selected a relevant discussion group for her information need. Adding narrative degrades performance slightly, which may be because of the addition of irrelevant or partly terms that broaden the scope of the query. These findings suggest that creating a comprehensive and effective topic representations that identify all the important relevance aspects in social book search information needs is not easy, even for trained annotators. Such topics often contain complex, multi-faceted aspects, which may be the reason why users turn to the forum in the first place, as current book search systems provide limited options to express complex needs.

We also experimented with reranking results by combining the retrieval score with a prior probability based on the Bayesian average of a book's LibraryThing ratings. These average ratings provide a reliable probability of relevance and lead to significant improvements in performance.

In future work we will look in more detail at the overlap and complementarity of the title and mediated query and the role of term frequencies in topic representations of the complex information needs in social book search. We will also study the role of the detailed narrative and experiment with extracting the most salient additional terms to improve the topic representations. One way would be to use parsimonious language models [2] to remove common conversational terms.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO projects # 612.066.513, 639.072.601, and 640.005.001) and by the European Communitys Seventh Framework Program (FP7 2007/2013, Grant Agreement 270404).

Bibliography

[1] F. Andriaans, M. Koolen, and J. Kamps. The importance of document ranking and user-generated content for faceted search and book suggestions. In S. Geva, J. Kamps, and R. Schenkel, editors, *Focused Retrieval of Content and Structure: 10th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2011)*, volume 7424 of *LNCS*. Springer, 2012.

[2] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings SIGIR 2004*, pages 178–185. ACM Press, New York NY, 2004.

[3] M. Koolen, J. Kamps, and G. Kazai. 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.

[4] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: a language-model based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, 2005.