Social Book Search: The Impact of Professional and User-Generated Content on Book Suggestions

Marijn Koolen[1] Jaap Kamps[1] Gabriella Kazai[2]

[1] University of Amsterdam, The Netherlands
[2] Microsoft Research, Cambridge UK,

ABSTRACT

The Web and social media give us access to a wealth of information, not only different in quantity but also in character—traditional descriptions from professionals are now supplemented with user generated content. This challenges modern search systems based on the classical model of topical relevance and ad hoc search. We compare classical IR with social book search in the context of the LibraryThing discussion forums where members ask for book suggestions. This paper is an compressed version of [2].

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process)*

General Terms: Experimentation, Measurement, Performance

Keywords: Book search, User-generated content, Evaluation

1. INTRODUCTION

The web gives access to a wealth of information that is different from traditional collections both in quantity and in character. Especially through social media, there is more subjective and opinionated data, which gives rise to different tasks where users are looking not only for facts but also views and interpretations, which may require different notions of relevance. In this paper we look at how search has changed by directly comparing classical IR and social search in the context of the LibraryThing (LT) discussion forums, where members ask for book suggestions. We use a large collection of book descriptions from Amazon and LT, which contain both professional metadata and user-generated content (UGC), and compare book suggestions on the forum with Mechanical Turk judgements on topical relevance and recommendation for evaluation of retrieval systems. Searchers not only consider the topical relevance of a book, but also care about how interesting, well-written, recent, fun, educational or popular it is. Such affective aspects may be mentioned in reviews, but Amazon, LT and many similar sites do not include UGC in the main search index. Our main research question is:

- How does social book search compare to traditional search tasks?

For this study, we set up the Social Search for Best Books (SB) task as part of the INEX 2011 Books and Social Search Track.[1] We want to find out whether the suggestions are complete and reliable enough for retrieval evaluation and how social book search is related to traditional search tasks. We also want to know if users prefer professional or UGC for judging topical relevance and for recommendation, and how standard IR models cope with UGC.

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

DIR 2013, April 26, 2013, Delft, The Netherlands.

2. SOCIAL SEARCH FOR BEST BOOKS

In this section we detail collection and the LT forum topics.

Collection The Amazon/LT collection [1] consists of 2.8 million book records from Amazon, identified by ISBN, extended with social metadata from LT, marked up in XML. These records contain title information, Dewey classification codes and Subject headings supplied by Amazon. The reviews and tags were limited to the first 50 reviews and 100 tags respectively during crawling. The professional metadata is more evenly distributed than the UGC. Books have a single classification code and most have one or two subject headings, although a small fraction has no professional metadata. Typical of UGC, popular books have many tags and reviews while many others have few or none. The median number of reviews and tags are 0 and 5 respectively. That is, the majority has no reviews but at least a handful of tags.

Topics LibraryThing users discuss their books in forums dedicated to certain topics. Many of the topic threads are started with a request from a member for interesting, fun new books to read. Other members often reply with links to works catalogued on LT, which we connected to books in our collection through their ISBN. These requests for recommendations are natural expressions of information needs for a large collection of online book records, and the book suggestions are human recommendations from members interested in the same topic. For the Social Search for Best Books task we selected a set of 211 topics, some focused on fiction and some on non-fiction books. For the Mechanical Turk experiment we focus on a subset of 24 topics.

MTurk Judgements We compare the LT forum suggestions against traditional judgements of topical relevance, as well as against recommendation judgements. We set up an experiment on Amazon Mechanical Turk to obtain judgements on document pools based on top-10 pooling of the 22 runs submitted by the 4 participating groups. We designed a task to ask Mechanical Turk workers to judge the relevance of 10 books for a given book request. Apart from a question on topical relevance, we also asked whether they would recommend a book to the requester and which part of the metadata—curated or user-generated—was more useful for determining the topical relevance and for recommendation. We included some quality assurance and control measure to deter spammers and sloppy workers. Averaged over workers the LT agreement is 0.52.

3. SYSTEM-CENTERED ANALYSIS

We compare system rankings of the 22 official runs based on the forum suggestions and on the MTurk relevance judgements. The Kendall's τ system ranking correlation between the forum sugges-

Table 1: MTurk and LT Forum evaluation (nDCG@10 and recall@1000) of runs over different index fields

	MTurk							
	Rel		Rec		Rel&Rec		LT-Sug	
Field	nDCG	recall	nDCG	recall	nDCG	recall	nDCG	recall
Title	0.212	0.601	0.260	0.545	0.172	0.591	0.055	0.350
Dewey	0.000	0.009	0.003	0.007	0.000	0.005	0.001	0.022
Subject	0.016	0.008	0.021	0.010	0.016	0.009	0.003	0.009
Review	**0.579**	0.720	**0.786**	**0.756**	**0.542**	**0.783**	**0.251**	**0.680**
Tag	0.368	0.694	0.435	0.665	0.320	0.718	0.216	0.602

tions for 211 topics and the MTurk judgements on the 24 topics is 0.36. This is not due to the difference between the 211 topics of the forum suggestions and the subset of 24 topics selected for MTurk, as the correlation between the forum suggestions of the 211 and 24 topic sets is $\tau = 0.90$. It could be that the forum suggestions are highly incomplete. Most topics have few suggestions (median is 7). If the suggestions are a small fraction of all relevant books, good and bad systems will perform poorly as the chances of ranking the few suggested books above other relevant books is small. However, the highest MRR score among the 22 runs is 0.481. This means that on average, over 211 topics, this system returns a suggested book in the top 2. If this only occurs for a few topics, it could be ascribed to mere coincidence, but over 211 topics, such a high average is unlikely due to chance. Based on this, we argue the forum suggestions are relatively complete but represent a different task from the ad hoc task modelled by the topical relevance judgements from MTurk. In [2] we also show that the forum suggestions behave differently from known-item topics.

Next, we created a number of our runs to compare the forum suggestions against the MTurk judgements. For indexing we use Indri, Language Model, with Krovetz stemming, stopword removal and default smoothing (Dirichlet, μ=2,500). The titles of the forum topics are used as queries. In our base index, each xml element is indexed in a separate field, to allow search on individual fields.

Generally, systems perform better on recommendation judgements (MTurk-Rec in Table 1) than on topical relevance judgments (MTurk-Rel), and their combination (MTurk-Rel&Rec) and worst on the forum suggestions (LT-Sug). The suggestions seem harder to retrieve than books that are topically relevant. The Title field is the most effective of the non-UGC fields. It gives better precision and recall than the Dewey and Subject fields across all sets of judgements. The Review field is more effective than the Tag field. Note that all runs use the same queries. Even though book titles alone provide little information about books, with the Title field the majority of the judged topically relevant books can be found in the top 1,000, but only a third of the suggestions. The review and tag fields have high R@1000 scores for all four sets of judgements. There is something about suggestions that goes beyond topical relevance, which the UGC fields are better able to capture. Furthermore, the retrieval system is a standard language model, which was developed to capture topical relevance. Apparently these models can also deal with other aspects of relevance. It also shows how ineffective book search systems are if they ignore reviews. Even though there are many short, vague and unhelpful reviews, there seems to be enough useful content to substantially improve retrieval. This is different from general web search, where low quality and spam documents need to be dealt with.

4. USER-CENTERED ANALYSIS

The MTurk workers answered questions on which part of the metadata is more useful to determine topical relevance and which part to determine whether to recommend a book. Workers could indicate the description does not have enough information to answer questions Q1 (topical relevance) and Q3 (recommendation). We see in Table 2 the fraction of books for which workers did not have enough information split over the descriptions with no reviews (column 2), at least one review (column 3), no tags (column 4) and at least 10 distinct tags (column 5). First, without reviews, workers indicate they do not have enough information to determine whether a book is topically relevant in 37% of the cases, and label the book as relevant in 30% of the cases. When there is at least one review, in only 1% of the cases do workers have too little information to determine topical relevance, but in 54% of the cases they label the book as relevant. Reviews contain important information for topical relevance. The presence of tags seems to have no effect, as the fractions are stable across books with different numbers of tags. We see a similar pattern for the recommendation question (Q3).

Table 2: Impact of presence of reviews and tags on judgements

		Reviews		Tags	
		0 rev.	≥1 rev.	0 tags	≥10 tags
Top. Rel. (Q1)	Not enough info.	0.37	0.01	0.09	0.09
	Relevant	0.30	0.54	0.49	0.48
Recommend. (Q3)	Not enough info.	0.53	0.01	0.14	0.12
	Rel. + Rec.	0.22	0.51	0.46	0.45

In summary, the presence of reviews is important for both topical relevance and recommendation, while the presence and quantity of tags plays almost no role.

5. CONCLUSIONS

In this paper we ventured into unknown territory by studying the domain of social book search with traditional metadata complemented by a wealth of user generated descriptions. We also focused on requests and recommendations that users post in real life based on the social recommendations of the forums. We observe that the forum suggestions are complete enough to be used as evaluation, but they are different in nature than traditional judgements for known-item, ad hoc and recommendation tasks. Even though most online book search systems ignore UGC, our experiments show that this content can improve both traditional ad hoc retrieval effectiveness and book suggestions and that standard language models seem to deal well with this type of data.

Our results highlight the relative importance of professional metadata and UGC, both for traditional known-item and ad hoc search as well as for book suggestions.

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 Community's Seventh Framework Program (FP7 2007/2013, Grant Agreement 270404).

REFERENCES

[1] T. Beckers, N. Fuhr, N. Pharo, R. Nordlie, and K. N. Fachry. Overview and Results of the INEX 2009 Interactive Track. In *ECDL*, volume 6273 of *LNCS*, pages 409–412. Springer, 2010.

[2] M. Koolen, J. Kamps, and G. Kazai. Social book search: Comparing topical relevance judgements and book suggestions for evaluation. In *Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM 2012)*. ACM Press, New York NY, 2012.