Evaluating Network-Aware Retrieval in Social Networks

Tom Crecelius
MPI Informatik, Saarbrücken, Germany
tcrecel@mpi-inf.mpg.de

Ralf Schenkel
Saarland University, Saarbrücken, Germany
schenkel@mmci.uni-saarland.de

ABSTRACT

This paper discusses the problem of evaluating search and recommendation methods in social tagging networks that make use of the network's social structure. While many such methods have recently been introduced, they fall short of evaluating the quality of the results they produce in a systematic way, which is mostly caused by the lack of publicly available test collections.

1. INTRODUCTION

Collaborative recommendations and search in social networks has been a very active research topic in recent years, and there has been an increasing number of papers proposing different methods and algorithms in this area. A recently upcoming trend is keyword-based search in social tagging networks such as del.icio.us, Flickr, or Librarything, where users annotate their items with tags. While early works in this area focused on frequency-based methods to evaluate searches, more recent approaches such as [3, 5, 6, 8] take the connections of the querying user in the social network into account when computing results. However, as there is neither a standard evaluation methodology nor a standard collection of data sets and topics, the proposals evaluate their techniques in different ways, making it impossible in practice to compare the performance of different techniques without reimplementing and reevaluating them. This clearly shows that there is an increasing demand for a publicly available evaluation platform to compare the performance of different methods for searching social tagging networks. This paper first discusses existing evaluation methods and demonstrates their shortcomings. It then proposes a future community-based evaluation task for this scenario.

2. EVALUATION APPROACHES

Evaluating effectiveness of search methods in social tagging networks has been recognized as an important yet unsolved problem [2]. A number of different evaluation methodologies for assessing the quality of such search methods has been proposed, typically as a byproduct of proposing a novel search method [4, 6, 8, 9]. Due to the lack of publicly available, large-scale samples of social networks, each paper uses a different data set, either crawls from the Web sites of exsting social networks (del.icio.us for [6], del.icio.us, Flickr and Librarything for [8, 9]) or non-public snapshots of such networks (del.icio.us for [3], data from inside IBM for [4]). As snapshots and crawls had been done at different instances in time, the crawls had used different techniques, and each snapshot had been only a small fraction of the whole network, it is very likely that each evaluation used a largely different data set, limiting the possibility to compare the results. While all approaches use reasonably large sets of keyword queries and some notion of average result precision to evaluate result quality, they drastically differ in how they determine the set of ground truth results.

User-Independent Ground Truth. Exploiting that del.icio.us maintains bookmarks, Bao et al [6] used the DMOZ catalogue of Web sites to extract queries and ground truth. They first selected DMOZ categories with enough urls that were also present as bookmarks in their del.icio.us crawl. For each such category, a keyword query was defined based on the category label. The set of relevant results for this query was formed by the urls in that category that were also present in the crawl. While this yields a large test collection, it completely ignores the user who submits the query. Methods that determine user-specific results are therefore potentially penalized by this method.

Context-based Ground Truth. Our own previous work [8] generated a set of relevant answers which we assumed to be more relative to the querying user. We computed the set of relevant answers for a keyword query as the set of items from friends of the querying user that conjunctively match the keyword query. However, this method comes with some bias towards network-aware search methods because it gives priority to results in the neighborhood of the user. Additionally, there is no clear evidence if those results really satisfy the user's information need. An item that does not appear among the user's friends may as well be very relevant for the user.

Temporal Ground Truth. If not a single snapshot, but a series of snapshots of the same social network is available, a set of relevant answers to a query can be defined by exploiting the change of the network over time. Given a tag query and a user, the set of relevant answers is formed by the items with (at least) these tags that the user aded in the future.[1] While such an item will definitely be important for the user (or she would not have added it to her collection), there is no guarantee that it is also relevant for this specific query. Additionally, an item may not have been added by the user simply because she didn't know about it, not because she found it irrelevant.

User Study. We performed a small-scale user study with six users in [9] that were actual users of the LibraryThing social network. The experiment was done in a Cranfield style, with a set of topics that each user defined, results for each topic from different methods pooled, and each pool assessed by the user who defined the topic. However, while the queries have been made public, the snapshot of the social network is not available, making it difficult to reuse them to evaluate other approaches.

[1] A temporal ground truth has been internally used by Yahoo! Research, but has not yet been published.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

3. A SETUP FOR A COMMUNITY-DRIVEN EVALUATION TASK

3.1 Collection

Data from *Bibsonomy*[2] is available for research purposes, and is currently being used for the ECML challenge [3]. However, this data set does not provide information about how users are connected, it is limited to the narrow domain of scientific publications, and it is of rather limited size, so it is unclear how significant results derived from this corpus could be. More interesting candidates for social networks would be large-scale networks like del.icio.us or librarything, which combine rich annotations and complex friend networks with interesting and rich content. However, it is unclear to which extent the owners of these networks would be willing to supply data from these networks.

Maintaining such a publicly available collection of – potentially sensitive – data from private networks raises some privacy issues. However, we think that these issues can be dealt with through a combination of technical and legal means: First, attempts should be made to anonymize the users contained in the snapshot, for example by assigning them unique, but random identifiers. As experiences with other collections, for example with the AOL query log [1] and the NetFlix dataset [7], have show in the past, such an anonymization cannot make sure that anonymous users cannot be mapped to their real identity. The collection should therefore be made available only under a restricted license that allows its use only for (possibly limited) research. This has been successfully done in the past for several other collections. Finally, the data in the collection can be restricted to information that is already available on the Web, thereby making it of limited use to anybody wanting to break the anonymity of users.

3.2 Users, Queries and Assessments

Community-driven evaluation venues such as INEX have been successfully distributing the load of defining queries and assessing evaluation results among the participating organizations. We propose to use a similar community-driven approach for the evaluation of search in social tagging networks. Each participating organization needs to define several possible *topics* (including a description of the information need, a corresponding keyword query, and example results). Each such topic must come with a *user* from that organization that is actually a user in the social network from which we take the data. In the ideal case, this would be a user who already has a long history of activity in this network, but it could as well be someone who joins for the experiment (and, of course, needs to upload and tag items and make connections to other users). Once topics are fixed, a snapshot of the network can be taken that includes these users (or, alternatively, a large crawl of the network could be performed using these users as crawl seeds).

Once the data set is available, participating organizations can—just like in existing benchmarks such as INEX or TREC—submit their results, which will then be pooled per topic and assessed by the original topic author. The latter is necessary because we assume that most topics will be of a highly subjective nature, with results highly depending on the submitting user, so we think that they cannot easily be assessed by someone who did not define the topic.

3.3 Primary Measurements

Evaluation measurements can be similar to those currently used for evaluating text retrieval methods. More specifically, there should be at least one precision-based metric such as NDCG, and one recall-based metric such as MAP.

3.4 Additional Measurements

Given that the evaluation process that we described so far incurs a great deal of work for all participants, an important subgoal of this activity would be to understand if the automatic methods for ground truth building described in Section 2 yield results that are comparable to the results with manual assessments. More precisely, it should be examined if the automated methods result in similar ranking of the different participating systems (not necessarily in similar absolute precision for the different runs). If that was the case for one of the methods, future evaluations could be much easier.

4. CONCLUSION

This paper introduced the problem of evaluating search methods in social tagging networks, presented several evaluation approaches used by different publications, and showed why none of them is generally applicable. We proposed a novel community-based evaluation that successfully captures the pecularities of social networks. However, the success of such an initiative eventually depends on the cooperation of the companies and institutions owning social network data, and on the agreement of enough organizations to participate in such a project.

5. REFERENCES

[1] E. Adar. User 4xxxxx9: Anonymizing query logs. In *Query Log Analysis: Social and Technological Challenges*, 2007.

[2] S. Amer-Yahia, M. Benedikt, and P. Bohannon. Challenges in searching online communities. *IEEE Data Eng. Bull.*, 30(2):23–31, 2007.

[3] S. Amer-Yahia, M. Benedikt, L. V. S. Lakshmanan, and J. Stoyanovich. Efficient network aware search in collaborative tagging sites. *Proc. VLDB Endowment*, 1(1), 2008.

[4] E. Amitay, D. Carmel, N. Har'El, S. Ofek-Koifman, A. Soffer, S. Yogev, and N. Golbandi. Social search and discovery using a unified approach. In *18th International World Wide Web Conference*, pages 1211–1211, April 2009.

[5] I. Assent. Actively building private recommender networks for evolving reliable relationships. In *M3SN Workshop*, pages 1611–1614, 2009.

[6] S. Bao, G.-R. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW*, 2007.

[7] S. Greengard. Privacy matters. *Commun. ACM*, 51(9):17–18, 2008.

[8] R. Schenkel, T. Crecelius, , M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, and G. Weikum. Efficient top-k querying over social-tagging networks. In *SIGIR*, 2008.

[9] R. Schenkel, T. Crecelius, M. Kacimi, T. Neumann, J. X. Parreira, M. Spaniol, and G. Weikum. Social wisdom for search and recommendation. *IEEE Data Engineering Bulletin*, 31(2):40–49, 2008.

[2] `http://www.bibsonomy.org`

[3] `http://www.kde.cs.uni-kassel.de/ws/dc09`