Bibliometrics in Online Book Discussions: Lessons for Supporting Complex Search Tasks

Marijn Koolen

Institute for Logic, Language and Computation, University of Amsterdam,The Netherlands

Abstract. Online book discussion forums provide rich information on how readers think about and describe books, how books are related to other books and how people search for and recommend books. Within the Social Book Search (SBS) Lab at CLEF we analyse book search requests on the LibraryThing forums and find several types of complex search tasks where bibliometrics naturally combines with information retrieval. This paper explores how book search information needs in online book discussions are related to bibliometric analysis and how the structure of book citations in reviews and discussions can support complex book search tasks where users want to go beyond topical relevance ranking and see how books are related to each other and particularly to the books and authors they know.

1 Introduction

Online book discussions are a rich source of information about books, readers and their preferences. There are several online platforms like GoodReads, LibraryThing and Reddit where people discuss the books they read and want to read. These discussions can be analysed using bibliometric techniques to get an insight in reading behaviour and interests and relationships between books and readers, which could potentially be exploited in retrieval systems to support book search tasks. Although bibliometrics is often associated with the domain of scholarly communication, its application in book-related social media can shed new light on the value of bibliometric information. For instance, the fact that anyone can cite or mention books in any context prompts questions about how this affects the meaning and value of citation counts and co-citation strengths and in their potential use in retrieval systems.

Book search in social media is investigated in the Social Book Search (SBS) Lab at CLEF 2015–2016, which provides IR test collections for a broad range of realistic and often complex search tasks, and includes several forms of user-generated content about books that can be analysed with bibliometric techniques. Book reviews by users, online discussion forums and user catalogues are different data sources providing connections between users, books and authors. Recommender systems typically exploit user interactions with items to derive interests and generate recommendation. But what the value of such interactions is for active searchers and interactive retrieval systems remains an open question.

This keynote explores the connections between the topics of bibliometrics in information retrieval[1] and social book search in two parts. The first part discusses book discussion and search information needs found in social media and how they reveal both a need for and resources for bibliometric analysis. The second part describes how bibliometric information can derived to support various complex search tasks.

2 Social Book Search

The Social Book Search (SBS) Lab investigates book search in social media, where users search for, discuss and review books, leaving rich connections between readers, books and authors. The SBS Lab started as a single system-centred evaluation track at INEX[2] in 2011 with a focus on studying the relative value of professional metadata and user-generated content for retrieval. Over the years, new types of professional and user data have been added and the focus has shifted on building systems for supporting the often complex search tasks that are found on the LibraryThing[3] (LT) discussion forums. These forums provide a great opportunity for studying realistic book search needs and recommendations as well as book mentions over time, in a variety of communities and in different contexts.

LT offers forum members a way to markup book and author mentions through so called *touchstones* to allow members to easily look up the mentioned works and authors and make explicit which author or work is discussed. Similar to wiki syntax, forum members can surround book titles and author names with brackets while typing their forum posts. The touchstone technology immediately links this to a specific work or author and allows the user to correct the identified work.

2.1 Social Book Search Requests

Users of the LT discussion forums often posts requests for book recommendations based on specific interests and information needs. Such requests have elaborate descriptions of what they like and have already read and what they are looking for. Many of the search requests encountered on the LT forums are complex. Searchers often have information needs that are vague or hard to describe, which they express through shortlists of books that represent what they are looking for. Or they describe sets of books that they are interested in and want to know what order to read them in, which of those books to start with or which to read next. Such tasks are not well-supported by current retrieval systems, which may be a reason they turn to the forum to ask for recommendation, and present opportunities for exploiting bibliometric information to aid search and retrieval. Bibliometric information may confer relevant wisdom from a niche of knowledgeable readers. Figure 1 shows the start of a discussion thread with the topic starter asking

[1] See BIR 2016 Workshop: http://www.gesis.org/en/events/events-archive/conferences/ecirworkshop2016/ and the Special issue "Combining Bibliometrics and Information Retrieval" http://link.springer.com/article/10.1007%2Fs11192-014-1484-3

[2] The INitiative for the Evaluation of XML Retrieval, see http://inex.mmci.uni-saarland.de/

[3] See: https://www.librarything.com/

Fig. 1. Book request on the LibraryThing forum

for recommendations on the topic of multiculturalism. The post contains touchstones to books and authors he is already familiar with, but also explains what he did not like about a specific book. This is a rich description of a highly complex search request.

2.2 Citations in User-Generated Content

Over the course of five years, several test collections have been developed with diverse document collections, including Amazon user reviews, LT user catalogues and LT forum discussion threads. The Amazon/LibraryThing (ALT) collection contains book descriptions from Amazon for 2.8 million books, including over 10 million Amazon user reviews and 320 million user tags from LT, and additional library records from the Library of Congress and the British Library. Next to the book descriptions, there is a set of over 94,000 user catalogues with 39 million catalogue entries of 5.6 million distinct works. The reviews, catalogues and discussion threads contain relations between books and users that can be analysed statistically and could be interpreted as polyrepresentations of citation and co-citation structure. But each of these 'citation' types has its own characteristics that introduce challenges in interpreting these structures.

Platforms like GoodReads and LibraryThing and online book shops like Amazon allow readers to review the books they have read. Zuccala and Bod [15] discuss how formal book reviews can be regarded as mega-citations and how they fit in citation theory. But in social media, reviews are often less formal, more heterogeneous in structure and content and can be written for many different reason, including to voice an opinion about the author or the topic of the book, its price or appearance. Although it is possible to use reviews by the same reviewer as a connection between books (i.e. as co-citations), there is not necessarily any meaningful relation in the fact that two books are reviewed by the same person, as they may have read them for different purposes or out of different interests.

Users on LT can create their personal catalogues of books they have read or want to read. These catalogues show which books a user added to her catalogue and when, as well as optional ratings and tags. Again, the books in a user catalogue are connected to each other through a user's diverse and changing interests, and often unknown reasons for cataloguing.

The LT discussion forums contain many groups and threads with topical focus, where books mentioned are related through the topic, but there are also many groups and threads based on reading challenges–e.g. reading 75 books in a year or 11 books in each of 11 categories in 2011–and games, where the relationships between mentioned books may be hard to interpret. The statistical structure in book mentions can be analysed at various levels: whole threads, individual posts, users or discussion groups. Discussion groups for instance may represent specific communities, such as the Science Fiction Fans group and the Military History group.

Citations and co-citations can be counted among all books that mentioned, or within meaningful subsets, e.g. all books classified under History & Geography, Social science or Literature & Fiction. A comparison of the reviews, catalogues and forum mentions reveals that fiction is more popular than non-fiction in terms of mentions (66%), but in terms of catalogued and reviewed books, non-fiction is the larger category (53% and 60% respectively). Among the non-fiction categories in the Dewey Decimal Classification, history and social science are the most prevalent categories. Among the most catalogued books, there are interesting differences in mention frequency. Among book series, the first one or two books are mentioned much more frequently than later books in the series. This is possibly because users want to give others quick access to information on books that these series start with. The rest of the titles are not as useful since people can easily find these further books through the first one. It may be that users avoid mentioning too obvious connections between books which could help systems avoid too obvious recommendations. Vice versa, among the books that are not frequently catalogued but mentioned relatively often there are mention nominated works for literary prizes such as the Man Booker Prize[4] and the Orange Prize.[5]

Bibliometric analysis also reveals interesting patterns in reading behaviour and information needs. Koolen et al. [5] categorised 944 forum book search requests into five groups: 1) known-item, 2) content-based, 3) familiarity-based, 4) content- & familiarity-based and 5) context-based. Content-based requests purely focus on the content of the book, e.g. topic, genre or plot. Familiarity-based requests describe previous reading experiences or specific books that the user knows and likes, with the information need being to find similar books or good follow ups. Context-based requests contain information needs that have to do with who the reader is (age, gender, background knowledge) and the context of reading (in class, in an airport). They found that for familiarity-based book search requests mostly represent fiction-related information needs, whereas requests for non-fiction are mainly content-based. Users who post content-based requests tend to have larger catalogues than users who post context-based and familiarity-based requests and get less popular books as suggestions.

[4] See: http://themanbookerprize.com/

[5] Now called the Baileys Women's Prize for Fiction, see http://www.womensprizeforfiction.co.uk/

2.3 Identifying Book Citations

Citations in scholarly communications are often explicit, but in informal discussions and reviews, works may be mentioned without explicit reference. This poses a problem for statistical analysis. In book-related discussions, citations can be explicit and specific or implicit, vague and generic. A forum post can mention an author's entire oeuvre without mentioning specific works, or refer to a specific work by description—a form of non-indexed eponymal citedness [10]—as in the following example: *"The second Chadbourn trilogy is ok, not as good as the first."*[6] This sentence mentions two trilogies by the author Mark Chadbourn, but neither of them by name. Such mentions are hard to detect automatically, but are a form of citation. This challenge is currently explored via the SBS 2016 Mining Track[7] which evaluates automatic detection and linking of book mentions.

3 Using Book Citations for Retrieval

A well-established way to use citations to improve retrieval effectiveness is use citations context. In scientific literature, the context in which articles are cited can improve retrieval effectiveness [7, 8]. What people other than the author say about an article enriches the representation. The same is true for books. Koolen et al. [3] looked at the effectiveness of including book discussions in the index as representations of a book. Both user tags and reviews are very effective representations for retrieval compared to book titles, library subject headings and other formal and curated metadata, with reviews being the most effective across a broad range of search tasks [2]. Although subject headings give very precise access, they are often either too specific or too general to cover an information need sufficiently. However, there are other ways in which different types of book citations can be useful for search and retrieval.

3.1 Search with shortlist

Searchers often base their information on previous reading experiences. On the LT forums, almost 36% of the identified requests explicitly mention the reading experience on which their need is based [5]. An example of such a request is "Can someone recommend a book that has all the joy, charm, numerous characters, pathos, adventure, love of language, etc. that the novel David Copperfield has?" (topic 10392[8]). Here, the previous reading experience is qualified with what aspects the searcher likes about the book. Searching by author is another form of search by shortlist that is encountered on the forums. In a set of 300 search requests on the LT forums, 15% were related to author names [4]. This suggests book readers want to get insight in how books are related to each other to determine which books are similar or complementary to what they have

[6] See message 41 in the following discussion thread: `http://www.librarything.com/topic/5990`

[7] See `http://social-book-search.humanities.uva.nl/#/mining`

[8] See: `http://www.librarything.com/topic/10392`

already read. One of the challenges for developing retrieval systems to support this is to establish what the relevant connection between shortlisted items is.

Searching with both a textual representation of the information need as well example books is related to entity ranking, specifically list completion tasks [1]. An important difference is that here the notion of list-membership is not objective and unambiguous. What is related to the given examples and of interest to the user is not necessarily clear to the user herself and can change while the user is gathering new information about books. Schnabel et al. [9] conducted user studies to test an interface that allowed users to construct shortlists of candidates films for consumption. This reduced the user's cognitive load in remembering possible candidates resulting in more exploration and higher satisfaction with the final selection. The information provided by shortlists could be used in retrieval as a form of query-by-document [14]. Finding similar items or users is typical in recommender systems, where the history of interactions represents a user's latent interests. But the book search requests on the LT forums show that users also actively search with these latent interests. This type of search seems to combine the retrieval and recommender paradigms, but is rarely supported by search and recommendation engines.

Query by one or more documents lends itself well for incorporating aspects of bibliometrics into the retrieval model. IR systems can exploit the statistical structure of book interactions to identify relevant relationships between documents and use these for retrieving and ranking, but can also show the user how books of interest are related to each other and to other books in the collection.

3.2 Reading order

Some book search requests have as underlying information need the best book to start reading in relatively clearly delineated selection of books, such as where to start with reading large oeuvres of prolific authors, or of specific series or specific sub-genres. These requests have a different aim than identifying what the relevant books are. The requester is typically aware of some or all of the books in a particular set, but wants to know if there is a natural or best order to read them in. In the case of series, there is often a natural order in which to read them, e.g. the order in which they were written or published, or the chronological order of the story. In many other cases it is not clear whether there is a most useful or interesting order and if so, what that order is. If the selection of books is based on subject area, readers may want to start with introductory texts, than explore more specific sub-topics based on their interests or needs. But not knowing the content of these books in advance, users may want guidance on selecting what to read next.

This is another search task where bibliometric information can enhance information retrieval. For instance, systems can show in what order other users have bought, catalogued or reviewed these books. Several signals can be derived from book citations that could potentially facilitate users in determining a useful reading order:

Popularity how many others have bought, catalogued or rated a book is a sign of where readers start in for instance the oeuvre of an author. Readers who read only a single book by an author probably start with the most well-known book. A big difference in

popularity of the books in a set may indicate that the most popular book is a good starting point for the user to see if she wants to read more.

Order of interaction: a useful reading order may be derived from the order in which other users catalogue, rate or review the books in a set. Although there may many differed subsets and orders in which users have done this, some sets may have a clearer 'natural' order than other sets. Using transitional probabilities between two books, a Markov model may reveal prevalent orders or paths through a set. To distinguish this from popularity, one could set a threshold to include only the interactions from users who have dealt with a minimum subset of the books under consideration. The order of interaction is to some extent necessarily related to publication order in that older books can be read before newer books are published.

(Co-)Citations: If a user already has a starting point and wants to know where to continue, co-citation information could be useful. For instance, Pennant diagrams ([12]—which show co-cited works of a seed work on a 2-dimensional plane with co-citation counts on the X-axis (TF) and inverse citation counts (IDF) on the Y-axis—have an interesting relation with specificity. White and Mayr [13] looked at the co-occurrence of subject descriptors, and argue that descriptors that occur infrequently but relatively often co-occur with a given seed descriptor, tend to be more topically specific in relation to that seed descriptor. For exploring books in a subject area, this may be exploited to help users identify the specificity of books within a subject.

To support these kinds of requests, the SBS Interactive Track investigates how a multi-stage interface offering different screens, each of which can support a specific phase of a complex search task. The underlying idea is derived from the information search process models of Kuhlthau [6] and Vakkari [11]. A browsing stage is offered to support the *pre-focus* phases in the model of Vakkari [11] where users have vague needs and want to explore the collection. A standard search stage with query and ranked results list supports the *focus* (Vakkari) or formulation (Kuhlthau) stage, and a book bag screen where selected books can be analysed and used for similarity searches (e.g. 'more like this') supports Vakkari's *post-focus* phase and Kuhlthau's collection and closure phases. Citation counts and contexts and co-citation information could be included in these stages, especially the browse and book bag stages, to support users in exploring the connections between the books and authors they know with the rest of the collection.

4 Conclusions

The domain of social book search offers search tasks and data where bibliometric techniques can be meaningfully incorporated in information retrieval systems. The complex search requests of the LibraryThing forum suggest users want to get insight in how books are related to each other and to previous reading experiences and in which order they should read them. This requires information about the book domain that text-based search cannot easily provide, but that bibliometric techniques naturally capture. These types of needs and types of relevance signals generalize beyond the book domain. Where to start reading in a subject area or genre is typical for researchers and students new to a field. Similar signals can be captured from reference management tools like Mendeley and Zotero, which are used to keep tracking of reading.

The informal nature of book citations on social media and the diversity of contexts and intentions pose specific challenges for getting a grip on bibliometric analysis and deriving useful knowledge from it, but there are social book data sets and IR test collections with which, through experimentation, the role and value of bibliometrics for information retrieval can be established.

References

[1] K. Balog, P. Serdyukov, and A. P. de Vries. Overview of the TREC 2011 entity track. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*. NIST, February 2012.

[2] M. Koolen. "User Reviews in the Search Index? That'll Never Work!". In M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann, editors, *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, volume 8416 of *Lecture Notes in Computer Science*, pages 323–334. Springer, 2014.

[3] M. Koolen, J. Kamps, and G. Kazai. Social book search: comparing topical relevance judgements and book suggestions for evaluation. In X. Chen, G. Lebanon, H. Wang, and M. J. Zaki, editors, *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 185–194. ACM, 2012.

[4] M. Koolen, G. Kazai, J. Kamps, M. Preminger, A. Doucet, and M. Landoni. Overview of the INEX 2012 social book search track. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, 2012.

[5] M. Koolen, T. Bogers, A. van den Bosch, and J. Kamps. Looking for books in social media: An analysis of complex search requests. In *ECIR 2015, Proceedings*, volume 9022 of *LNCS*, pages 184–196, 2015. ISBN 978-3-319-16353-6.

[6] C. C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science*, 42(5):361, 1991.

[7] A. Ritchie, S. Robertson, and S. Teufel. Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008*, pages 213–222. ACM, 2008. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458113. URL http://doi.acm.org/10.1145/1458082.1458113.

[8] A. Ritchie, S. Teufel, and S. Robertson. Using terms from citations for IR: some first results. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008*, volume 4956 of *Lecture Notes in Computer Science*, pages 211–221. Springer, 2008. ISBN 978-3-540-78645-0.

[9] T. Schnabel, P. N. Bennett, S. T. Dumais, and T. Joachims. Using shortlists to support decision making and improve recommender system performance. *CoRR*, abs/1510.07545, 2015. URL http://arxiv.org/abs/1510.07545.

[10] E. Száva-Kováts. Non-indexed indirect-collective citedness (niicc). *Journal of the American Society for Information Science*, 49(5):477–481, 1998. ISSN 1097-4571. doi: 10.1002/(SICI)1097-4571(19980415)49:5⟨477::AID-ASI9⟩3.0.CO;2-8. URL http://dx.doi.org/10.1002/(SICI)1097-4571(19980415)49:5<477::AID-ASI9>3.0.CO;2-8.

[11] P. Vakkari. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of documentation*, 57(1):44–60, 2001.

[12] H. D. White. Combining bibliometrics, information retrieval, and relevance theory, part 1: First examples of a synthesis. *Journal of the Association for Information Science and*

Technology, 58(4):536–559, 2007. doi: 10.1002/asi.20543. URL http://dx.doi.org/10.1002/asi.20543.

[13] H. D. White and P. Mayr. Pennants for descriptors. *CoRR*, abs/1310.3808, 2013. URL http://arxiv.org/abs/1310.3808.

[14] Y. Yang, N. Bansal, W. Dakka, P. Ipeirotis, N. Koudas, and D. Papadias. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 34–43, New York, NY, USA, 2009. ACM.

[15] A. Zuccala and R. Bod. Book reviews as 'mega-citations': A fresh look at citation theory. In *Proceedings In 17th international conference on science and technology indicators, STI 2012*, 2012.