Social Book Search: Comparing Topical Relevance Judgements and Book Suggestions for Evaluation

Marijn Koolen[1] Jaap Kamps[1,2] Gabriella Kazai[3]

[1] Archives and Information Studies, University of Amsterdam, The Netherlands
[2] ISLA, Informatics Institute, University of Amsterdam, The Netherlands
{marijn.koolen,kamps}@uva.nl
[3] Microsoft Research, Cambridge UK,
v-gabkaz@microsoft.com

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: How does their effectiveness transfer to the changing nature of information and to the changing types of information needs and search tasks? We use the INEX 2011 Books and Social Search Track's collection of book descriptions from Amazon and social cataloguing site LibraryThing. We compare classical IR with social book search in the context of the LibraryThing discussion forums where members ask for book suggestions. Specifically, we compare book suggestions on the forum with Mechanical Turk judgements on topical relevance and recommendation, both the judgements directly and their resulting evaluation of retrieval systems. First, the book suggestions on the forum are a complete enough set of relevance judgements for system evaluation. Second, topical relevance judgements result in a different system ranking from evaluation based on the forum suggestions. Although it is an important aspect for social book search, topical relevance is not sufficient for evaluation. Third, professional metadata alone is often not enough to determine the topical relevance of a book. User reviews provide a better signal for topical relevance. Fourth, user-generated content is more effective for social book search than professional metadata. Based on our findings, we propose an experimental evaluation that better reflects the complexities of social book search.

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 has made the landscape of search more complex. Traditional IR models were developed in a time when the information that was available was limited. Retrieval systems indexed titles, abstracts and keywords assigned by professional cataloguers for collections of officially published documents. On the web, there is much more information. Every aspect of human life is published on the web, which leads to different search tasks and different notions of relevance. Traditional IR was mainly based on the ad hoc search methodology of a user who wants information that is topically relevant to her information need [24]. Many state-of-the-art retrieval systems are still based on classical IR models and are evaluated using this ad hoc search methodology. Increasingly, research in areas such as web [10], blog [18] and realtime search [26] has focused on new search tasks in this changing environment.

In this paper we aim to study how search has changed by directly comparing classical IR and social search. Sites like Amazon and LibraryThing offer an opportunity to do this, as they provide traditional descriptions—titles, abstract and keywords of books—as well as user-generated content data in the form of user tags, reviews, ratings and discussions. To gain more insight in these changes, we compare classical IR with social book search in the context of the LibraryThing discussion forums, where members ask for book suggestions. We use a large collection of book descriptions from Amazon and LibraryThing, which contain both professional metadata and user-generated content, and compare book suggestions on the forum with Mechanical Turk judgements on topical relevance and recommendation for evaluation of retrieval systems. Amazon and LibraryThing are typical examples were users can add their own content about books, but like many similar sites, do not include user-generated content in the main search index. Any direct searching in the collection is done on professional metadata. One reason for users to ask for suggestions on discussion forums may be that they cannot search sites directly on the subjective content provided by other users, which indicates these suggestion are more than just topical relevance judgements.

Relevance in book search—as in many other scenarios—is a many-faceted concept. There may be dozens or hundreds of books that are topically relevant, but the user often wants to know which one or two to choose. This is where the information need goes beyond topical relevance: searchers also care about how interesting, well-written, recent, fun, educational or popular it is. Some of these facets are covered by professional metadata, such as subject headings for topical facets, and publication data for recency, size, binding and price. Affective aspects, such as how well-written and interesting a book is, is not covered by professional metadata, but can be covered by reviews. Social book search has elements of subject search as well as recommendation. We use the book requests and suggestions as a real world scenario of book search, and as examples of relevance judgements, with the aim to investigate how this search task differs from traditional ad hoc search. Can we em-

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$15.00.

ulate these scenarios with known-item search or traditional ad hoc retrieval based on topical relevance? We use Amazon Mechanical Turk to obtain judgements about the topical relevance of books as well as about recommendation. 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] One of the goals of this track is to build test collections for this and other book search tasks. The book requests from the forum are used as information needs and the book suggestions as relevance judgements. These are real information needs and human suggestions. With these suggestions we avoid problems with pooling bias [5]. We hope to find out whether the suggestions really are the best books on the topic or just a sample of a much larger set of books that are just as good. The latter case would mean the list of suggested books is incomplete. We compare these suggestions with judgements of topical relevance and recommendation, which we obtained through Amazon Mechanical Turk[2] (MTURK). Specifically, we address the following questions:

- Can we use book requests and suggestions from the LibraryThing forum as topics and relevance judgements for system evaluation?
- How is social book search related to known-item search, ad hoc search and recommendation?
- Do users prefer professional or user-generated content for judging topical relevance and for recommendation?

Professional metadata is evenly distributed—no single book is privileged. A book usually has only one classification number, and often no more than two or three subject headings. For user-generated content this is dramatically different. The amount of content added is related to how many users added content, which leads to a more skewed distribution. Popular books may have many more ratings, reviews and tags than less popular books. This leads to the following questions:

- How do standard IR models cope with user-generated content?
- How effective are professional and user-generated content for book suggestion?

The rest of this paper is organized as follows. We first discuss related work in Section 2. Next, we describe the search task and scenario in detail in Section 3, and then describe the document collection, information needs and the Mechanical Turk experiment in Section 4. We discuss the system-centered evaluation in Section 5, and the user-centered evaluation in Section 6. Finally, we draw conclusions in Section 7.

2. RELATED WORK

In this section, we discuss related work on novel search tasks, classical information retrieval based on controlled vocabularies, and crowdsourcing in IR.

2.1 Search Tasks

At TREC, many of the evaluations still focus on the ad hoc search methodology where the aim is to find information that is topically relevant. Other evaluations have addressed that change in search task caused by a change in the information environment. There is information on the web of any level of subjectivity and quality. Research areas such as web search [10] and blog search [18] have identified search tasks very different from traditional subject search in catalogues, where other aspects of relevance play a role. For web search aspects of popularity and authority [19] and diversity [6] are important, for blog and twitter search, aspects of subjectivity [18] and credibility [26] play a role. [22] interviewed 194 book readers about their reading experiences and book selections. She found that readers welcome recommendations from known and "trusted" sources to reduce the number of candidates for selection and like to know what other readers have chosen. Reading a book is a substantial investment of time and energy, so searchers use a variety of clues to choose one or a few books from among a much longer list. This is supported by [21], who identified 46 factors that influenced children's assessment of relevance when selecting books along dimensions such as content, accessibility, engagement and familiarity.

2.2 Controlled Vocabularies and Retrieval

The Cranfield tests for evaluating information retrieval systems [7] showed that indexing based on natural language terms from documents was at least as effective for retrieval as formal indexing schemes with controlled languages. However, controlled vocabularies still hold the potential to improve completeness and accuracy of search results by providing consistent and rigorous index terms and ways to deal with synonymy and homonymy [14, 23]. One of the problems with traditional metadata based on controlled vocabularies and classification schemes is that it is difficult for both indexers and searchers to use properly. On top of that, searchers and indexers might use different terms because they have different perspectives. Buckland [4] describes the differences between vocabularies of authors, cataloguers, searchers, queries as well as the vocabulary of syndetic structure. With all these vocabularies used in a single process, there is the possibility of mismatch. Users of library catalogues use keyword search, which often does not match the appropriate subject headings [2, p.7]. People use the principle of least effort in information seeking behavior: they prefer information that is easy to find, even if they know it is of poor quality, over high quality information that is harder to find. [2, p.4] One of the interesting aspects of user-generated metadata in this respect is that it has a smaller gap with the vocabulary of searchers [17].

Tags have also been compared to subject headings for book descriptions with the growing popularity of sites like Delicious, Flickr, and LibraryThing. Tags can be seen as personal descriptors for organizing information. Golder and Huberman [8] distinguish between tags based on their organizing functions. What (or who) it is about, what it is, who owns it, refining categories, qualities or characteristics, self reference and task organizing. Lu et al. [16] compared LibraryThing tags and LCSH. They find that social tags can improve accessibility to library collections. Yi and Chan [28] explored the possibility of mapping user tags from folksonomies to Library of Congress subject headings (LCSH). They find that with word matching, they can link two-thirds of all tags to LC subject headings. In subsequent work [27], they use semantic similarity between tags and subject headings to automatically apply subject headings to tagged resources.

Peters et al. [20] look at the retrieval effectiveness of tags taking into account the tag frequency. They found that the tags with the highest frequency are the most effective. Kazai and Milic-Frayling [12] incorporate social approval votes based on external resources for searching in a large digitized book corpus. They evaluate their model with a set of queries from a book search transaction log and traditional topical relevance judgements by paid assessors. Their

[1] https://inex.mmci.uni-saarland.de/tracks/books/
[2] http://www.mturk.com

results show that social approval votes can improve a BM25F baseline that indexes both full-text and MARC records.

2.3 Crowdsourcing Relevance Judgements

There is a lot of recent research on using crowdsourcing for relevance assessment [1, 9]. To make sure the quality of judgements is sufficient, numerous quality-control measures have been proposed [11, 13, 15]. A minimal approval rate (how many of the previous tasks have been approved by the task owner), trap questions ("check this box if you did NOT read the instructions"), captcha's and flow-dependent questions (the next question depends on the answer to the previous question) are all effective quality-control mechanisms. Crowdsourced relevance judgements have been effectively used at INEX to evaluate book page retrieval tasks [13].

3. SOCIAL SEARCH FOR BEST BOOKS

In this section we detail the Social Search for Best Books (SB) task as run at INEX 2011, and the used collection.

3.1 Social Book Search Task

The goal of the SB task is to evaluate the relative value of controlled book metadata versus user-generated or social metadata for retrieving the most relevant books for search requests on online book discussion forums. Controlled metadata, such as the Library of Congress Classification and Subject Headings, is rigorously curated by experts in librarianship. On the other hand, user-generated content lacks vocabulary control by design. However, such metadata is contributed directly by the users and may better reflect the terminology of everyday searchers. Both types of metadata seem to have advantages and disadvantages. With this task we want to investigate the nature of book search in an environment where book descriptions are a mixture of both types of metadata, with the aim to develop systems that can deal with more complex information needs and data sources.

The scenario is that of a user turning to Amazon Books and LibraryThing to search for books they want to read, buy or add to their personal catalogue. Both services host large collaborative book catalogues that may be used to locate books of interest. On LibraryThing, users can catalogue the books they read, manually index them by assigning tags, and write reviews for others to read. Users can also post messages on a discussion forum asking for help in finding new, fun, interesting, or relevant books to read. The forums allow users to tap into the collective bibliographic knowledge of hundreds of thousands of book enthusiasts. On Amazon, users can read and write book reviews and browse to similar books based on links such as "customers who bought this book also bought... ". Neither service includes reviews or tags in the search index. Users have to browse through individual book descriptions to be able to search through the user-generated content.

The SB task assumes a user issues a request to a retrieval system, which returns a (ranked) list of book records as results. This request can be a list of keywords, a natural language statement. We assume the user inspects the results list starting from the top and works her way down until she has either satisfied her information need or gives up. The retrieval system is expected to order results by relevance to the user's information need. User requests can be complex mixtures of topical aspects ("I want a book about X"), genre aspects (fiction/non-fiction, poetry, reference), style aspects (objective/subjective, engaging, easy-to-read, funny), and other aspects such as comprehensiveness, recency, etc. The user context, i.e., their background knowledge and familiarity with specific books, adds further complexity. They might have found a number of books already, read some of them and discarded other options, and want to know what else is available. This aspect of user context was left out of the SB task in the first year but will be included in future years. Participants of the SB task are provided with a set of book search requests from the LibraryThing discussion forums and are asked to submit the results returned by their systems as ranked lists. We assume one of the reasons why readers turn to the discussion forums is that they can ask such complex questions that are hard to address with current search engines.

Table 1: Statistics on the Amazon/LibraryThing collection

type	*min*	*max*	*median*	*mean*	*std. dev.*
Professional					
Dewey	0	1	1	0.61	0.49
Subject	0	29	1	0.66	0.72
BrowseNode	0	213	18	19.84	10.21
User-generated					
Tag	0	50	5	11.45	14.55
Rating/Review	0	100	0	5.05	14.98
Automatic					
Similar product	0	15	1	2.37	2.40

3.2 Professional and User Generated Book Information

To study social book search, we need a large collection of book records that contains professional metadata and user generated content, for a set of books that is representative of what readers are searching for. We use the INEX Amazon/LibraryThing corpus [3].

The collection consists of 2.8 million book records from Amazon, extended with social metadata from LibraryThing, marked up in XML.[3] This set contains books that are available through Amazon. These records contain title information as well as a Dewey Decimal Classification (DDC) code and category and subject information supplied by Amazon. Each book is identified by its ISBN. Since different editions of the same work have different ISBNs, there can be multiple records for a single intellectual work. Each book record is an XML file with fields like <isbn>, <title>, <author>, <publisher>, <dimensions>, <numberofpage> and <publicationdate>. Curated metadata comes in the form of a Dewey Decimal Classification in the <dewey> field, Amazon subject headings are stored in the <subject> field, and Amazon category labels can be found in the <browseNode> fields. The social metadata from Amazon and LibraryThing is stored in the <tag>, <rating>, and <review> fields. The reviews and tags were cut-off after the first 50 reviews and 100 tags respectively during crawling.

How many of the book records have curated metadata? In the Amazon/LibraryThing data, there is a DDC code for 61% of the collection and 57% has at least one subject heading. The classification codes and subject headings together cover 78% of the collection. There is also a large hierarchical structure of categories called *browseNodes*, which is the category structure used by Amazon. All but 296 books in the collection have at least one browseNode category. Most records have just one Dewey code and one subject heading (Table 1), while some records have no Dewey code or subject heading. Records never have more than one Dewey code (to determine the location of the physical book on the shelves), but can have multiple subject headings. The low standard deviation of the subject headings indicates that the distribution is flat. The BrowseNode distribution is more skewed, with a median (mean) of

[3] See `https://inex.mmci.uni-saarland.de/data/nd-agreement` for information on how to get access to this collection.

18 (19.84) BrowseNode categories, but a minimum of 0 and a maximum 213. The median number of subject headings per book is 1. For the next edition of this task at INEX we extend the collection with records from the British Library and the Library of Congress, which may have more headings per book.

How many of the book records have user-generated metadata? Just over 82% of the collection has at least one LibraryThing tag, but less than half (47%) has at least one rating and review. The median (mean) number of tags per record is 5 (11.45) and the median (mean) number of ratings and reviews is 0 (5.05). The distribution of the amount of UGC is thus much more skewed than the distribution of the amount of professional metadata. This is due to a popularity effect. Multiple users can add content to a book description, and popular books will receive more tags and reviews than less popular books. This is an important difference between professional and user-generated content. UGC not only lacks vocabulary control, but also introduces an imbalance in the exhaustivity and redundancy of book descriptions. The impact of this imbalance is discussed in Sections 5 and 6.

4. SOCIAL BOOK RECOMMENDATIONS

In this section we describe the book recommendation requests at the LibraryThing (LT) forums, and the Mechanical Turk experiment we ran to obtain relevance judgements.

4.1 Topics and Recommendations

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. They describe what they are looking for, give examples of what they like and do not like, indicate which books they already know and ask other members for recommendations. Other members often reply with links to works catalogued on LT, which have direct links to the corresponding records on Amazon. 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. Each topic consists of a title, group name, thread, narrative and so-called ‘touchstones’.

Title of the topic, a short description of what the topic is about.

Group name identifying the discussion group where the topic was posted.

Narrative describing the topic, it is the first message in the thread explaining what the topic creator is looking for.

Thread containing the messages posted by members of the discussion group in response to the initial request.

Touchstones the list of books suggested by the members, identified by LT work ID. Members can use a Wiki-type syntax around the title of a work to have LT automatically identify it as a book title and link it to the a dedicated LT page on that book. When LT misidentifies a book, members can and often do correct the link.

We distributed the topics, which included the Title, Group name and Narrative to participants of the INEX 2011 Book Track, who could use any combination of these fields for retrieval. We note that the title and narrative of a topic may be different from what the user would submit as queries to a book search system such as Amazon, LT, or a traditional library catalogue. However, as the message is addressed to other members, we consider this a natural expression of the information need. As an example, consider a topic titled *Help: WWII pacific subs* from a user in the *Second World War History* discussion group, with the following narrative:

> *Can anyone recommend a good strategic level study of us sub campaign in pacific? All I seem to scare up is exploits of individual subs. I have ordered clay blairs big study but I would like something from this decade if it exists.*

The topic of the request is strategy of the US submarines in the Pacific in World War 2. The user has done some searching and has found books on US submarines, but no strategic studies. Furthermore, the user wants something recent and something good. The latter qualification is subjective. Does the user mean comprehensive or accurate, easy to read or engaging, or all of these? The user already knows about and ordered a relevant book by Clay Blair. The thread has eight replies in which five books are recommended and automatically identified in the Touchstone list, including the one by Clay Blair that the topic creator already ordered.

We note that the suggestions are made by other forum members than the requester, and the requester may consider only few or even none as interesting enough. However, we argue that these suggestions are valuable judgements that are relevant to the information need, because they are made by members of the same discussion group. We assume they share this topical interest with the requester and suggest books they have read or know about.

We use these suggested books as initial relevance judgements for evaluation. Some of these suggestions link to a different book from the one intended, and suggested books may not always be what the topic creator asked for, but merely be mentioned as a negative example or for some other reason. From this it is clear that the collected list of suggested books can contain false positives and is probably incomplete as not all relevant books will be suggested (false negatives), so may not be appropriate for reliable evaluation. The suggestions as relevance judgements avoid the problem of pooling bias [5]. Although the judgements were pooled by a number of LT members, these LT members are not evaluated.

We crawled over 18,000 topics from the forums, with over 11,000 topics having at least one suggested book. We filtered these using regular expressions such as “I’m looking for” and “can you recommend” and a number of others to locate topics that have actual book requests. This resulted in 1,800 topics, from which we manually selected all topics that really contain a request for book suggestions, reducing the set to 945 topics. The other topics contained requests ranging from information from non-book sources, tips on how to do something or places to go to related to their topic. We use the titles of the topic threads as natural succinct expressions of the information need. Many of these 945 titles do not reflect the actual information needs, which would make them unsuitable as queries. We ran all 945 titles as queries on a full-text index of our collection (see Section 5.2 for indexing details) and kept only those topics for which at least 50% of the books suggested by the forum members were retrieved, leaving us with 211 topics from 122 discussion groups. We note that this introduces a bias towards topics for which the full-text index gets high recall. However, we think that the other topics would introduce noise in the evaluation and creating our own queries for them would reduce the realistic nature of the topic set. The 211 topics form the official topic set for the Social Search for Best Books task in the INEX 2011 Book Track. For the Mechanical Turk experiment we focus on a subset of 24 topics.

We manually classified topics as requesting fiction or non-fiction books, or both, as there are some topics where the creator requested both fiction and non-fiction books. In total, there 79 fiction topics (37%), 122 non-fiction topics (58%) and 10 mixed topics (5%). For our selection of 24 topics, we selected 12 fiction and 12 non-fiction topics. Arguably, fiction-related needs are less concerned with the topic of a book than non-fiction needs, and more with genre, style and affective aspects like interestingness and familiarity. For such needs it seems more clear that the traditional IR approach of gathering topical relevance judgements is the wrong task model.

4.2 MTurk Judgements

We want to 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-k pooling.

The SB task had 4 participating teams who together submitted 22 runs. From the 211 topics in the total set, we manually selected 24 topics with a short and clear request for which to obtain relevance judgement from MTURK. The books to be judged are based on top 10 pools of all 22 official runs. In cases where the top 10 pools contained fewer than 100 books, we increased the pool depth to the smallest rank k at which the pool contained at least 100 books.

We designed a HIT (Human Intelligence 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. At the beginning of the HIT we asked how familiar they are with the topic and afterwards how difficult the HIT was, which they could answer with a 5-point Likert scale.

As on Amazon, we show only the 3 most helpful reviews. Each review has a total number of votes T and a number of helpful votes H with $H \leq T$. On Amazon, the most helpful review seems to be determined by the number of helpful votes and the ratio of helpful to total votes. We use $ln(H+1)*(\frac{H}{T})^n$ to score helpfulness, where n controls the relative weight of the ratio $\frac{H}{T}$. With $n = 3$ we found the resulting ranking of reviews to closely resemble the ranking of the top 3 reviews for books on Amazon. For popular books with many reviews and votes, we expect the votes to filter out bad reviews and review spam (fraudulent reviews written to promote or damage a book, author or publisher). For more obscure books with few or no votes, helpfulness has little impact and fake reviews may be selected. It is not clear how many fake reviews there are, how to identify them, nor what their impact is. We therefore do not address this issue in this paper.

We asked the following questions per book:

Q1. Is this book useful for the topic of the request?
Workers could pick one of the following answers

- Very useful (perfectly on-topic).
- Useful (related but not completely the right topic).
- Not useful (not the right topic)
- Not enough information to determine.

Q2. Which type of information is more useful to answer Q1?
Workers see a 5-point Likert scale, with *Official description* on the left side and *User-generated description* on the right side.

Q3. Would you recommend this book?
Workers could pick one of the following answers:

- Yes, this is a great book on the requested topic.
- Yes, it's not exactly on the right topic, but it's a great book.
- Yes, it's not on the requested topic, but it's great for someone interested in the topic of the book.
- No, there are much better books on the same topic.
- I don't know, there is not enough information to make a good recommendation (skip Q4).

Q4. Which type of information is more useful to answer Q3?
Again, workers could choose on a five-point scale between *Official description* and *User-generated description*.

Q5. Please type the most useful tag (in your opinion) from the LibraryThing tags in the User-generated description, with a text box and next to it a check box with the text *(or tick here if there are no tags for this book.)*

In addition, workers could give optional comments in a comment box per book. We included some quality assurance and control measure to deter spammers and sloppy workers, and approved new assignments once a day over a period of 6 days.

LT agreement Each HIT contained at least one book that was recommended on the LT forums. Workers doing multiple HITs can easily be checked on agreement with LT forum members. For workers who do only one or two HITs, agreement cannot be reliably determined and is not used for approval. Once workers did 3 or more HITs, we rejected a HIT if it made their LT agreement level drop below 60%.

Relevance contradiction A worker first saying a book is related, then saying it is on-topic is inconsistent, but is not contradicting her- or himself. We consider the answers to Q1 and Q3 to be contradicting when a worker answers *on-topic* for Q1, then *unrelated* for Q3 or the other way around. Also, when a worker answers *not enough information* for Q1, then either *on-topic*, *related* or *unrelated* for Q3.

Type contradiction A metadata type contradiction is made when a worker answer that the UGC is more useful than the professional metadata when there is no UGC.

Tag occurs Finally, we asked workers to type in the most useful tag from the UGC (or tick the adjacent box when the UGC contains no tags). The LibraryThing tags were placed at the bottom of the UGC description, so this question forced workers to at least scroll down to the bottom of the description and check if there are tags.

Qualification Based on previous MTurk experiments, we used two worker qualifications. Workers had to have an approval rate of 95% with at least 50 approved HITs—i.e. only workers whose previous work on MTurk was of high quality—and we only accepted workers registered in the US.

We created a total of 272 distinct HITs. With 3 workers per HIT we ended up with 816 assignments. Only 7 assignments were rejected, either because workers skipped the last few books in the HIT (4 cases) or because their agreement was too low (3 cases).

In total, there were 133 different workers, of which 90 did only one HIT, 13 did two HITs and 30 workers did three or more. The distribution of HITs per worker is highly skewed, with more than half of the 816 HITs done by only 7 workers. This power-law-like distribution is typical of crowdsourcing experiments [1, 13]. Averaged over workers the LT agreement is 0.52. Low agreement was found for workers who did only one HIT, where there is only one data point to compute agreement, which is not enough to reliably compute agreement or reject a HIT. Workers who did at least 3 HITs (covering 86% of all HITs) have a median (mean) LT agreement of 0.67 (0.65). Averaged over assignments the agreement is

0.84, which shows that the few workers who did many HITs scored very high on agreement.

There are only 18 Relevance contradictions, spread over 15 approved HITs. From these, we discarded the books with contradicting judgements. No Type contradictions were made. In the answer categories of both the topical relevance and recommendation questions, we used the same levels of topical relevance (*perfectly on-topic, related, unrelated*). If workers choose the same level of topical relevance for both Q1 and Q3, or *not recommended* or *not enough information* for Q3, their answers are consistent, which was the case for 95% of the assignments. Time to complete a single HIT ranged between 3 and 111 minutes with an average of 13 minutes and 9 seconds. These numbers suggest workers performed most HITs conscientiously. Per Worker, an average of 68% of the tags they filled in for Q5 exactly matched a tag in the book description (median 70%). When there was no matching tag, this was mostly because workers combined two separate tags or made misspellings.

Most workers are not very familiar with the search topics for which they have to judge books. On a scale from 0 (totally unfamiliar) to 4 (very familiar), the median (mean) familiarity is 1 (1.5). For 3 topics the median familiarity is 0, for 12 topics it is 1, for 8 topics it is 2 and for 1 topic it is 3. Although workers are not very familiar with the topic of request, they indicate the work is not difficult. On a scale from 0 (very easy) to 4 (Very difficult), for 21 topics the median difficulty is 1 (fairly easy) and for 3 topics the median difficulty is 2 (medium difficulty). For only 9 assignments (1%) workers thought the HIT was very difficult, for 86 assignments (11%) they chose 3 (fairly difficult). We discuss the results of the MTurk experiment in the user-centered analysis in Section 6.

5. SYSTEM-CENTERED ANALYSIS

In this section we focus on system-centred evaluation. We want to know whether the forum suggestions are similar to any of the three known tasks—known-item search, ad hoc search, and recommendation—and whether the suggestions are complete and reliable enough for evaluation. First, we look at the official submissions of the Social Search for Best Books task, and compare the system rankings of the different sets of judgements. Second, we use additional runs we created ourselves to compare different index fields for professional metadata and user-generated content.

5.1 Comparing System Rankings

If we want know whether two sets of relevance judgements can be used to evaluate the same retrieval task, we can compare the system rankings they produce. If the sets of judgements model the same task, they should give the same answer when asked to choose which of two systems is the better one. We compare the system rankings of the 22 officially submitted runs based on the topical relevance judgements from MTurk and on the LT forum suggestions. We use Kendall's Tau and Tau$_{AP}$ [29]. The latter puts more weight on ranking the top scoring systems similarly than on ranking the lower scoring systems similarly.

The set of relevance judgements based on the suggestions for the 211 forum topics is denoted as LT-211, the subset of 24 topics selected for MTurk, but still using the forum suggestions as relevance judgements is denoted as LT-24 and with the Amazon MTurk topical relevance judgements as AMT-24-Rel. The system rank correlations are shown in Table 2. Recall that the subset of 24 topics is not randomly selected. The LT-24 subset still leads to a similar system ranking as the LT-211 set. The forum suggestions seem robust against non-random selection. The system ranking based on the AMT-24-Rel judgements is very different from those of the forum suggestions. The difference between τ and τ_{AP} is bigger between the AMT-24-Rel judgements and the two LT sets, showing that mainly disagree on the top systems.

Table 2: Kendall's τ and τ_{AP} system ranking correlations on nDCG@10 between the three sets of judgements (τ/τ_{AP})

	LT-24	AMT-24-Rel
LT-211	0.90/0.83	0.39/0.20
LT-24	–	0.36/0.19

Why do these sets produce such different system rankings? The AMT-24-Rel judgements are based on the top 10 results of all the official submissions, so the nDCG@10 scores do not suffer from incomplete judgements. The LT forum suggestions are not based on pools, but are provided by a small number of forum members who may have limited knowledge of all the relevant books. It could be that their suggestions are highly incomplete, and that many of the top 10 results of the official runs are just as relevant.

To get a better idea of the completeness of the forum suggestions we zoom in on the best scoring runs (the top one being a Language Model run that uses all user-generated content and pseudo-relevance feedback). The best system has a Mean Reciprocal Rank (MRR) of 0.481 and a Precision at rank 10 (P@10). of 0.207. There are several systems from different participants that get lower but similar scores. Considering that most topics have a small number of suggestions (the median number of suggested books is 7), these are remarkably high scores, and indicate the system is performing well. In a collection of millions of books, this retrieval system picks out several of the small number of books suggested by forum members. This indicates that the suggestions by forum members are not an arbitrary sample of a much larger set of books that are relevant to the topic, but are a relatively complete set in and of themselves. If the suggestions were only a small sample from a set of equally relevant books (say 7 out of 100, thus highly incomplete), the chances of a retrieval system consistently (for 211 topics) ranking at least one of those 7 at rank 2 or 3 are very small. The suggestions form a set of books that stand out. With top-k pooling the above argument cannot be made, since the small number of judgements is biased towards the evaluated systems. But this is not a pooling effect, since the suggestions are independent of the submitted runs. With a P@10 of 0.207, the best performing system ranks 2 of the suggested books, out of a collection of 2.8 million, in the top 10, on average over 211 topics, lending further support that the suggestions are relatively complete.

5.2 Effectiveness of Metadata Fields

For indexing we use Indri,[4] Language Model (without belief operators), 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. For the LibraryThing tags we create two versions of the index. One where we index distinct tags only once (Tag Set) and one where we use the tag frequency (how many users tagged a book with the same tag) as the term frequency (Tag Bag). That is, if 20 users applied tag t to book b, the Tag Set index will have a term frequency of 1 for (b, t) and the Tag Bag index will have a term frequency of 20 for (b, t).

The book records have unique ISBNs, but some records are different editions of the same *intellectual work*. Having multiple versions of the same work in the ranking is redundant for the user, so we ignore any other version after the first version found in the rank-

[4] URL: `http://lemurproject.org/indri/`

Table 3: Known-item and forum suggestion evaluation of runs over different index fields

	Known-item			Forum suggestions		
Field	MRR	R@10	R@1000	MRR	R@10	R@1000
Title	0.414	0.540	**0.820**	0.118	0.048	0.350
BrowseNode	0.004	0.000	0.240	0.083	0.028	0.261
Dewey	0.000	0.000	0.000	0.002	0.000	0.022
Subject	0.010	0.020	0.020	0.012	0.002	0.009
Review	**0.480**	**0.680**	0.800	**0.382**	**0.227**	**0.680**
Tag (set)	0.118	0.220	0.540	0.213	0.125	0.616
Tag (bag)	0.227	0.400	0.560	0.342	0.178	0.602

ing. To identify multiple manifestations of the same work, we use the mappings provided by LibraryThing.[5] With these mappings, we replace the ISBNs in the result lists and in the judgements with LibraryThing work IDs. With duplicate IDs in the ranking we keep only the highest ranked result with that ID.

5.2.1 Known-item versus Forum Suggestions

It is possible that the small set of suggestions are ranked high because book suggestion is very similar to known-item search. To check this possibility, we created a set of 50 known-item topics. We pooled all the suggested books for all 211 topics and randomly selected 50 books, to make sure the known-item topics target books from the same distribution.

There is a popularity effect that can explain why reviews and tag frequency work well. There is a plausible overlap between the people who buy, tag and review, e.g., *historical fiction* books and the people who suggest books in the *historical fiction* groups. Their suggestions are probably based on the books they have read, which are the books that they made popular. Is our finding a trivial one then? Not at all. They could suggest very different books from the ones that every historical fiction fan reads, or could be a non-representative sample of historical fiction readers.

The Known-item evaluation results of the individual metadata fields are shown in Table 3. The Title field is very effective. The controlled subject access fields are not at all effective, which is not surprising since they serve a different purpose. The tags are more effective than the controlled subject access points, but less than the title. The reviews are the most effective field, even outperforming the title field. Named access points in the formal metadata are effective for known-item search, but user-generated content without any formal and controlled metadata can be just as effective.

The competitiveness of the Title field for known-item topics is in stark contrast with its low scores for the forum suggestions. Book suggestion on the LT forum seems different from known-item search. Next, we compare the book suggestions with traditional topical relevance judgements.

5.2.2 MTurk evaluation results

The performance of systems on the topical relevance judgements (AMT-Rel) is shown in columns 2–4 in Table 4, on the recommendation judgements (AMT-Rec) in columns 5–7 and on the topical relevance + recommendation (AMT-Rel&Rec) in columns 8–10. The results for the forum suggestions (LT-Sug) are in columns 11–13. Generally, systems perform better on AMT-Rec than on AMT-Rel, and AMT-Rel&Rec and worst on LT-Sug. The suggestions seem harder to retrieve than books that are topically relevant. The exception is that the Review field is more effective for AMT-Rel&Rec than for topical relevance alone, apart form nDCG@10. Reviews become more effective when there is a recommendation element involved. The Title field is the most effective of the non-UGC fields. It achieves better precision and recall than the BrowseNode, Dewey and Subject fields across all sets of judgements. The Dewey and Subject fields are the least effective fields. The Review field is more effective than the Tag field. The bag of tags is more efficient than the set of tags for precision, but less effective for recall. The review and tag fields have similar R@1000 for all four sets of judgements. This last observation merits further discussion. The title field is reasonably effective for the AMT judgements, which are based on judgement pools from the 22 official submissions, which used much more than just the title field. The Title field scores between 0.601 for R@1000 (recall at rank 1,000) for topical relevance, but 0.35 for the forum suggestions. Note that for all runs and sets of judgements, the queries are the same. Even though book titles alone provide little information about books, with 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. 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.

The official submissions all used UGC, creating a bias in the judgement pools. The runs based on professional metadata have a larger fraction of non-judged results in the top ranks than the runs based on UGC. The performance of the Title field on the AMT judgements may be an underestimation. This cannot be the case for the LT forum suggestions, as they have no pool bias.

It also suggests the workers find the reviews more useful for topical relevance and recommendation than any other part of the book descriptions. Note that the LT forum members may not have seen any of the Amazon reviews before they made suggestions, whereas the workers were explicitly pointed at them, which could at least partly explain the higher scores for the AMT judgements.

We also looked at the difference between fiction-related requests and requests for non-fiction books. There are no meaningful differences between the two topic types. All runs score slightly better on the fiction topics, by the same degree, which is probably due to the fact that fiction topics have more suggested books than non-fiction topics. The different types of metadata have the same utility for forum suggestions of fiction and non-fiction topics.

It may not seem surprising that the longer descriptions of the reviews are more effective than the shorter descriptions of the other metadata fields. What is surprising however, is 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.

6. USER-CENTERED ANALYSIS

In this section we compare the MTurk judgements with the book suggestions from a user perspective, extending the analysis of system effectiveness above. The 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. On top of that, we can also look at the relation between the amount of user-generated content that is available and the particular answer given. As mentioned before, the amount of user-generated content is more skewed than the amount of professional metadata.

[5] http://www.librarything.com/feeds/thingISBN.xml.gz

Table 4: MTurk and LT Forum evaluation of runs over different index fields

	AMT-Rel			AMT-Rec			AMT-Rel&Rec			LT-Sug		
Field	nDCG10	MAP	R@1000	nDCG10	MAP	R@1000	nDCG10	MAP	R@1000	nDCG10	MAP	R@1000
Title (field)	0.212	0.105	0.601	0.260	0.107	0.545	0.172	0.088	0.591	0.055	0.040	0.350
BrowseNode	0.096	0.052	0.322	0.142	0.056	0.321	0.083	0.046	0.328	0.043	0.031	0.261
Dewey	0.000	0.000	0.009	0.003	0.000	0.007	0.000	0.000	0.005	0.001	0.001	0.022
Subject	0.016	0.002	0.008	0.021	0.002	0.010	0.016	0.003	0.009	0.003	0.002	0.009
Review	**0.579**	**0.309**	0.720	**0.786**	**0.389**	**0.756**	**0.542**	**0.333**	**0.783**	**0.251**	**0.174**	**0.680**
Tag (set)	0.337	0.173	**0.744**	0.422	0.199	0.711	0.288	0.158	0.754	0.125	0.097	0.616
Tag (bag)	0.368	0.182	0.694	0.435	0.197	0.665	0.320	0.176	0.718	0.216	0.154	0.602

Table 5: 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

6.1 Overlap between LT and MTurk

What is the overlap between the books suggested by forum members and the books judged by workers? Recall that we added at least one forum suggestion to each HIT. Of the 8,260 answers, 1,516 are for books that were suggested on the forums. Workers labelled 47% of all books as topically relevant (23% as related to the topic of request). In contrast, they labelled 66% of the suggested books as topically relevant (a further 18% at least related). Topical relevance is an important aspect for suggestions. For the recommendation question, 43% of all books are labelled as relevant and recommended (15% as related and recommended), and 62% of suggested books (13% related and recommended). If we consider only the recommendation aspect, 69% of all books and 80% of suggested books are recommended. Books suggested on the forum are more often recommended than other topically relevant books.

6.2 Relevance, Recommendation and UGC

How do forum suggestions compare with MTurk labels in terms of the amount of UGC? Recall that workers could indicate the description does not have enough information to answer questions Q1 (topical relevance) and Q3 (recommendation). Is this answer related to the number of reviews and tags in a description? We see in Table 5 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. With no tags, workers have too little information to determine topical relevance in 9% of the cases and label a book as relevant in 49% of the cases, and with at least 10 tags this is 9% and 48% respectively. The percentages are the same for books with at least 40 or 50 tags.

We see a similar pattern for the recommendation question (Q3). When there is no review, workers find it difficult to make a recommendation–not enough information in 53% of the cases, and only in 22% of the cases do they recommend a book. With at least one review, there is not enough information in only 1% of the cases, and a book is recommended in 51% of the cases. As with topical relevance, the presence and number of tags has little impact on recommendation. Without tags there is not enough information for recommendation in 14% of the cases and 46% of the books are recommended. With at least 10 tags, there is not enough information for 12% of the books and 45% is recommended.

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. It seems they do not provide user with additional value on top of the professional metadata, even though tags are more effective for retrieval in terms of topical relevance and recommendation (see Table 4). As with the system-centered analysis, we split the data over fiction and non-fiction topics but observed no difference. Workers seem to use the same metadata for requests for fiction books and requests for non-fiction books.

Some workers provided comments to explain their judgements. The following comments for the topic on recent books about US submarines in the pacific illustrate how user-generated content affects judgements:

- **Not enough information**:
  "Couldn't do much with no information but a title.",
  "I have a title that states submarines but that isn't enough.",

- **Related**:
  "This is fiction, and I think the person was asking for reference."
  "I'd be worried about recommending this one. It was described by users as being rather subjective."

- **Relevant, not recommended**:
  "Again, no description on the book, but going by the title, this might also work for the requester.",

- **Relevant + recommended**:
  "The user-generated review was so enthusiastic, I would recommend it just based on that. A memoir is still fiction-y but could be useful."
  "Looks good, and from 2001. So far, this would be my main recommendation choice."

The first comments indicate how professional metadata is often not specific enough. A novel on submarines in the pacific is considered not relevant because it is fiction. One worker does not recommend a book about submarines because it is "rather subjective" while another recommends a memoir because the review is so enthusiastic. These comments reveal the complexity of relevance in book search.

Table 6: Impact of the presence of reviews on metadata preference

	Q2. Relevance			Q4. Recommendation		
	all	0 rev.	≥1 rev.	all	0 rev.	≥1 rev.
Prof.	0.29	0.51	0.20	0.16	0.33	0.10
Equal	0.27	0.40	0.21	0.18	0.22	0.17
UGC	0.43	0.06	0.57	0.53	0.08	0.71
Skip	0.02	0.02	0.02	0.12	0.37	0.03

We assume most workers have not read any or only a few of the books they judge. Without having read the book, professional metadata and tags are not sufficient to determine whether a book is relevant or to make a recommendation. When there are reviews, workers almost always have enough information to determine relevance and make a recommendation. However, workers seem to use only one review, which may be an efficiency aspect. They get paid a fixed amount per HIT, so they can earn more per time unit by reading fewer reviews per book. They have no incentive to read more reviews, because to them the recommendation has little value.

Do users consider UGC as more of the same content or as content of a different nature? Tags seem to provide information of a similar nature to professional metadata. Reviews on the other hand radically affect the judgement of workers. Although workers seem to use only one review, the presence of reviews makes it easy for workers to make a recommendation and also helps in determining the topical relevance of books.

6.3 Utility of Metadata Types

To determine the relevance of books, do users prefer professional metadata, or UGC, or are they equally happy with either? If they prefer UGC, is this because it provides more metadata than the curated metadata? Or because it provides a different kind of metadata? Tags are similar in nature to subject headings [25, 28], while ratings and reviews are more opinionated and evaluative.

The distribution of preferences for metadata types is given in Table 6. In column 2 we see the fraction of all answers for each type for topical relevance. The professional metadata is considered more useful to judge the topical relevance of books in 29% of the cases, equally useful to UGC in 27% of the cases and less useful in 43% of the cases. The UGC is on average more useful than the professional metadata. For recommendation (column 5), UGC is considered more useful in the majority of cases (53%), while in only 16% of the cases the professional metadata is considered more useful. For recommendation, the number of cases where the question is skipped is given is much higher (12%) than for topical relevance (2%), which is mainly when workers indicated the book description does not provide enough data to make a recommendation, in which case they were asked to skip Q4.

How is the preference for professional metadata or UGC related to presence of reviews? The relation between the presence of at least one review and the utility of metadata types for topical relevance is shown in Table 6, in columns 3 and 4. The difference between no reviews and at least one review is big. With no reviews, most workers find professional metadata more useful for topical relevance, but 40% of workers find the two types of metadata equally useful. Only 6% find UGC more useful. With at least one review, this completely changes. The majority of workers finds UGC more useful and only 20% find professional metadata more useful. We found that further reviews do no affect the distribution, which suggests again that workers only use one review, even when multiple reviews are available.

For recommendation (columns 6 and 7 in Table 6), we see a similar pattern in the relation between the number of reviews and the utility of metadata types. With no reviews, the utility of UGC is low, but for recommendation, the lack of reviews makes it harder to answer the question; in more than a third of the cases (37%), workers skipped the question. This is strongly related to the answer given to Q3 (*Would you recommend this book?*). In 88% of the cases where the question is skipped, workers indicated at Q3 that there was not enough information to make a recommendation. When there is not enough information for recommendation, the question which type of metadata is more useful is hard to answer sensibly. This gives further evidence of workers filling in the questions seriously. When there is at least one review, the number of skipped questions drops to 3% and for 71% of the cases workers found the UGC more useful. Not surprisingly, UGC is even more important for recommendation than for determining topical relevance.

7. CONCLUSIONS

In this paper we ventured into unknown territory by studying the domain of book search that has rich descriptions in terms of traditional metadata descriptions—structured fields written by professionals—now complemented by a wealth of user generated descriptions—uncontrolled tags and reviews from the public at large. We also focused on the actual types of requests and recommendations that users post in real life based on the social recommendations of the forums. Relevance in book search—as in many other scenarios—is a many-faceted concept. Searchers do not only care about topical relevance (sometimes not at all), but also about how interesting, well-written, recent, fun, educational or popular it is.

We expected the forum suggestions, based on the collective knowledge of those answering the request, to cover only a small sample of the potentially relevant books. If this were the case, systems would perform poorly when evaluated on these suggestions, due to large numbers of retrieved and potentially relevant but unjudged documents. High precision is hard to achieve for a system that did not contribute to the pool of judged documents, if the judged relevant documents are highly incomplete. A system could still get a high precision for a single topic by accident. However, over 211 topics, a high precision is improbable. Yet, a standard IR model using an index based on user-generated content scores high on MRR and nDCG@10, even with a small number of suggested books in a collection of millions of book records. Hence, we observe that the forum suggestions are complete enough to be used as evaluation. The system ranking over all 211 topics correlates strongly with that of a non-random subset of 24 topics. This approach to test collection building based on forum requests and suggestions models a realistic, modern search task, that seems robust against topic selection and avoids pooling bias.

Next, we wanted to know how social book search is related to standard tasks like known-item search, ad hoc search on topical relevance, and topical recommendation. The system rankings of official submissions on the forum suggestions have a low correlation with those based on topical relevance judgements. Experiment with our own indexes also indicate suggestion is different. Book titles and professional metadata are both effective for known-item search, book titles give decent recall on topical relevance tasks, but neither is effective for the forum suggestions. However, part of the poor performance on the MTURK judgements may be due to a pooling bias. In contrast, user-generated content is much more effective for all tasks, including book suggestion. The LT forum suggestions seem different in nature than known-item topics and the MTURK judgements on topical relevance and recommendation.

Standard language models seem to deal well with the skewed distribution of user-generated content across book descriptions. The low effectiveness of professional metadata may also be partly due to a lack of useful term frequency information within a book description. However, the short book titles perform much better on topical relevance than on forum suggestions, indicating this is mainly a problem for aspects of relevance other than topicality. Although we have not explored all possible ways to exploit professional metadata, the user-centred evaluation corroborates our finding that user-generated content is more effective than professional metadata and covers more than topical relevance.

Even though most online book search systems ignore user-generated content, our experiments show that this content can improve traditional ad hoc retrieval effectiveness and is essential for book suggestions.

In the final part of our investigation we looked at how MTURK workers valued professional and user-generated content. The amount of tags has little impact on how useful they are for workers, and may perform similar functions to professional metadata. Workers on MTURK find reviews more useful than professional metadata and user tags, both for topical relevance and recommendation. For recommendation it seems obvious that ratings and opinionated reviews are more useful than objective tags and subject headings. For topical relevance, it may be that reviews contain more detail to determine how a book bears on the information need behind a book request on the LT forums.

How is social book search related to traditional search tasks? Topical relevance is a necessary condition of book suggestion, but not a sufficient one. Not all topically relevant books are suggested or recommended, indicating that other (more subjective) aspects also play a role. These other aspects are better captured by user-generated content than by professional metadata: reviews are more useful for the book suggestions on the forums. In future work we will incorporate profiles and personal catalogue data from forum members which may help capturing the affective aspects of book search relevance. Our results highlight the relative importance of professional metadata and user-generated content, 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] O. Alonso and R. A. Baeza-Yates. Design and Implementation of Relevance Assessments Using Crowdsourcing. In *ECIR 2011*, volume 6611 of *LNCS*, pages 153–164. Springer, 2011.

[2] M. J. Bates. Task Force Recommendation 2.3 Research and Design Review: Improving user access to library catalog and portal information. In *Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium*, 2003.

[3] 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.

[4] M. Buckland. Vocabulary as a Central Concept in Library and Information Science. In *Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. CoLIS3*, 1999.

[5] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Inf. Retr.*, 10(6):491–508, 2007.

[6] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, 2008. ACM.

[7] C. W. Cleverdon. The Cranfield tests on index language devices. *Aslib*, 19:173–192, 1967.

[8] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[9] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *CSLDAMT '10*, pages 172–179, 2010.

[10] D. Hawking and N. Craswell. Very large scale retrieval and web search. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. MIT Press, 2005.

[11] G. Kazai. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *ECIR 2011*, volume 6611 of *LNCS*, pages 165–176. Springer, 2011.

[12] G. Kazai and N. Milic-Frayling. Effects of Social Approval Votes on Search Performance. *Information Technology: New Generations, Third International Conference on*, 0:1554–1559, 2009.

[13] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for Book Search Evaluation: Impact of HIT Design on Comparative System Ranking. In *SIGIR*. ACM Press, New York NY, 2011.

[14] F. W. Lancaster. *Vocabulary control for information retrieval*. Information Resources Press, Arlington VA, second edition, 1986.

[15] J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 21–26, 2010.

[16] C. Lu, P. Jung-ran, and X. Hu. User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science*, 36(6): 763–779, 2010.

[17] A. Mathes. Folksonomies - Cooperative Classification and Communication Through Shared Metadata, December 2004.

[18] I. Ounis, C. Macdonald, M. de Rijke, G. Mishne, and I. Soboroff. Overview of the TREC 2006 Blog Track. In *TREC*, volume Special Publication 500-272. NIST, 2006.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[20] I. Peters, L. Schumann, J. Terliesner, and W. G. Stock. Retrieval Effectiveness of Tagging Systems. In *Proceedings of the 74rd ASIS&T Annual Meeting*, volume 48, 2011.

[21] K. Reuter. Assessing aesthetic relevance: Children's book selection in a digital library. *JASIST*, 58(12):1745–1763, 2007.

[22] C. S. Ross. Finding without seeking: the information encounter in the context of reading for pleasure. *Information Processing & Management*, 35(6):783 – 799, 1999.

[23] E. Svenonius. Unanswered questions in the design of controlled vocabularies. *JASIS*, 37(5):331–340, 1986.

[24] E. M. Voorhees. The Philosophy of Information Retrieval Evaluation. In *CLEF '01*, pages 355–370, 2002. Springer-Verlag.

[25] J. Voss. Tagging, folksonomy & co - renaissance of manual indexing? *CoRR*, abs/cs/0701072, 2007.

[26] W. Weerkamp and M. de Rijke. Credibility improves topical blog post retrieval. In *Proceedings of ACL-08: HLT*, pages 923–931, June 2008. Association for Computational Linguistics.

[27] K. Yi. A semantic similarity approach to predicting Library of Congress Subject Headings for social tags. *JASIST*, 61(8): 1658–1672, 2010.

[28] K. Yi and L. M. Chan. Linking folksonomy to Library of Congress Subject Headings: an exploratory study. *Journal of Documentation*, 65(6):872–900, 2009.

[29] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR*, pages 587–594. ACM, 2008.