An In-Depth Analysis of Tags and Controlled Metadata for Book Search

Toine Bogers[1], Vivien Petras[2]
[1]Department of Communication & Psychology, Aalborg University Copenhagen, Copenhagen, Denmark
[2]Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Berlin, Germany

Abstract
Book search for information needs that go beyond standard bibliographic data is far from a solved problem. Such complex information needs often cover a combination of different aspects, such as specific genres or plot elements, engagement or novelty. By design, subject information in controlled vocabularies is not always adequate in covering such complex needs, and social tags have been proposed as an alternative. In this paper we present a large-scale empirical comparison and in-depth analysis of the value of controlled vocabularies and tags for book retrieval using a test collection of over 2 million book records and over 330 real-world book information needs. We find that while tags and controlled vocabulary terms provide complementary performance, tags perform better overall. However, this is not due to a popularity effect; instead, tags are better at matching the language of regular users. Finally, we perform a detailed failure analysis and show, using tags and controlled vocabulary terms, that some request types are inherently more difficult to solve than others.

Keywords: book search, controlled vocabularies, social tagging, query analysis, failure analysis
Citation: Bogers, T., & Petras, V. (2017). An In-Depth Analysis of Tags and Controlled Metadata for Book Search. In iConference 2017 Proceedings, Vol. 2 (pp. 15-30). https://doi.org/10.9776/17004

Contact: toine@hum.aau.dk, vivien.petras@ibi.hu-berlin.de

1 Introduction

To locate a scroll in the ancient library of Alexandria, searchers were required to know the genre and author name (Phillips, 2010). Asking a librarian or other scholars at the library was probably the preferred strategy for searchers then. The process of locating relevant or interesting books to read has become substantially easier over the years with the advent of sophisticated book recommender systems, such as those offered by Amazon, which uses purchase history to suggest interesting books, as well as book search engines, such as Google Books which allows for full-text search through millions of books[1].

However, full-text search is not always a practical option. Book metadata is often the only way of satisfying book search requests. Previous work has compared the benefits of different types of metadata elements, such as bibliographic metadata, controlled vocabulary[2] (CV) terms, and user-generated content (e.g., reviews and tags) and found that user-generated content is the most effective in retrieving relevant books (Koolen (2014); (Bogers & Petras, 2015)). A second focus of our earlier work in (Bogers & Petras, 2015) was on the specific comparison between tags and CV terms. Tags were found to outperform CV terms significantly overall, but it was not a clear-cut advantage: some book search requests were better served by CV. Both metadata elements appear to provide complementary performance. We ascribed the overall advantage of tags to the popularity effect of tags—identical tags are repeatedly used by different users to describe the same book.

However, many unanswered questions remain: Is there really a true popularity effect for tags? Which types of book search requests are better addressed using tags and which using CV? Which book search requests fail completely for both metadata sources, and what characterizes such requests? In this paper, we provide a deeper understanding of the value of tags and CV for book retrieval by revisiting and updating our study, resulting in the following contributions:

1. A comparative analysis of tags and CV, focusing on complementarity and potential popularity effects.
2. A detailed analysis of book search requests that shows which types of information needs work better with tags or CV.

[1]http://www.newyorker.com/business/currency/what-ever-happened-to-google-books, last visited November 27, 2016.
[2]In this paper, we use the term *controlled vocabulary* to denote any form of taxonomy, categorization or language-controlled terminology (e.g., subject headings) that prescribes the form or term for a certain concept that is described (Dextre Clarke, 2008).

3. A failure analysis to determine why certain book search requests succeed while others fail.

The structure of this paper is as follows. We start in Section 2 with an overview of the relevant related work. Section 3 describes the experimental methodology used in this study. Section 4 explains the results of our comparative analysis of tags and CV for fulfilling book search requests. Section 5 contains the results of our request analysis showing which requests are better fulfilled by tags or CV. Section 6 describes our analysis of successful and failed requests. Finally, Section 7 discusses the outcomes of this study and concludes with suggestions for future work.

2 Background

In this section, we briefly discuss the relevant related research on (1) book search and information needs, (2) retrieval using CV versus user-generated content, and (3) aspects of search request analysis.

2.1 Book Search and Information Needs

While full-text search of books has been relatively underrepresented in research (Willis & Efron, 2013), book search in library catalogs has received plenty of attention recently (Slone, 2000; Kim, Feild, & Cartright, 2012; Saarinen & Vakkari, 2013). Magdy and Darwish (2008) showed that for shorter queries using only metadata can be just as effective as full-text retrieval, highlighting the importance of our work.

The Social Book Search (SBS) workshops[3] have been a fertile ground for research on book search since their inception in 2011, using metadata from Amazon, LibraryThing and library catalogs. The 2016 edition (Koolen et al., 2016) focused on the entire process of book search, from automatically detecting and categorizing book search requests to improving retrieval algorithms to investigating how people interact with book search interfaces.

The book search requests used in the SBS workshops were collected from the LibraryThing forums and represent complex, real-world information needs that are typically much longer and richer than the short queries submitted to conventional book search engines like Amazon or Google Books. Koolen, Bogers, Van den Bosch, and Kamps (2015) performed a detailed analysis of these book search requests and found different relevance aspects that went beyond the information present in traditional book metadata, such as novelty, engagement, and familiarity. Mikkonen and Vakkari (2016) found most book search requests for fiction were centered around familiarity and bibliographic information and an earlier study by Buchanan and McKay (2011) found that book requests are rooted in cultural context. Our study contributes to this body of work by providing a detailed analysis of which type of book search requests might be more effectively served by tags or by CV.

2.2 User-Generated Content vs. Controlled Vocabularies

The debate on the relative merits and drawbacks of controlled vocabularies versus free-text (including user-generated content) has been and continues to be a fruitful subject for small-scale case studies, with disagreement on which source is more effective (Cleverdon & Mills, 1963; Rowley, 1994; Dextre Clarke & Vernau, 2016). Recent, larger-scale work using the SBS collection has shown that user-generated content allows for more effective retrieval, with reviews being especially beneficial (Koolen (2014); (Bogers & Petras, 2015)).

Our more specific comparison of tags and CV in (Bogers & Petras, 2015) showed that while retrieval using tags yielded better results, CV and tags were successful for different search requests, a result also found in previous studies on retrieval using natural language vs. CVs (Gross & Taylor, 2005). We provide an in-depth analysis of the reasons behind these findings.

One possible reason for the complementary effects on retrieval performance might be the different characteristics of tags and CV. CV in the form of classes or categories from library classifications or even the Amazon taxonomy are very broad. Studies categorizing LibraryThing tags have found them to contain more subjective, contextual, and personal descriptions (Lawson, 2009; Voorbij, 2012), whereas subject headings tend to be more abstract and are required to be objective, impersonal, and only cover the most important topics of a book (LoC Cataloging Policy and Support Office, 2016).

[3] See `http://social-book-search.humanities.uva.nl\/#\/overview` (last accessed September 11, 2016) or Koolen et al. (2013).

While some studies concluded that tags and subject headings are complementary (Smith, 2007; Bartley, 2009), other studies found that tags either cover the same topics as subject headings or simply provide a more expansive terminology for book search (Heymann & Garcia-Molina, 2009; Lu, Park, & Hu, 2010). These studies also demonstrated that tags contain self-referential terms, which might introduce noise for a search engine, and might not cover less popular books equally well.

2.3 Search Request Analysis

Research on search requests or query analysis has taken on three forms: (1) query classification, (2) failure analysis, and (3) difficulty prediction. Query classification focuses on determining which types of information needs users of a particular information system have. For instance, Koolen et al. (2015) classified a set of LibraryThing forum requests used in the SBS labs as well as in this study.

Failure analysis looks at why requests fail to retrieve relevant results on particular collections. For instance, a thorough failure analysis of TREC queries found that the reasons for failure are usually due to semantic relationships represented in the query that are not understood by the search system (Buckley, 2009).

Finally, difficulty prediction involves studying how difficult it will be to retrieve relevant documents for a specific search request. Analyses distinguish between *pre-retrieval methods*, based mostly on linguistic features of the requests, and *post-retrieval methods*, which analyze requests with respect to documents in the collection or retrieved set. A good summary for these approaches can be found in Carmel and Yom-Tov (2010).

In this study, we perform a failure analysis using some pre-retrieval difficulty prediction indicators such as request or document length to research which book search requests can be more effectively fulfilled by tags or CVs and why some requests are bound to fail.

3 Methodology

This section describes the book metadata that was searched, the book search requests, relevance assessments and evaluation measures we used for the analyses.

3.1 The Amazon/LibraryThing Book Collection

The Amazon/LibraryThing collection has been used for several years in the Social Book Search workshops[2]. It was collected by Beckers, Fuhr, Pharo, Nordlie, and Fachry (2010) and contains over 2.8 million book records aggregated from Amazon, the British Library (BL), the Library of Congress (LoC), and LibraryThing (LT). Book records (henceforth referred to as 'documents') consist of over 40 different metadata elements, including core bibliographic metadata such as author or title, which were found to benefit retrieval (Bogers & Petras, 2015). In this paper, we only focus on the relative benefits of tags and CV; for an experimental comparison of the other metadata elements we refer the reader to Bogers and Petras (2015).

To put tags and CV on a more equal experimental footing and to be able to examine how they compare for individual documents, we filtered the original Amazon/LT collection so that all book records that did not contain at least one CV term *and* at least one tag were removed. This resulted in a test collection with 2,060,758 documents.

The tags in the Amazon/LT collection were originally collected from LibraryThing. They make up our Tags test collection[4]. The CV terms come from three different providers: Amazon, BL, and LoC. We combined these sources into a single test collection called CV . Table 1 shows the metadata elements making up the CV and Tags test collections.

[4]A test collection is the entire collection of over 2 million filtered book records used for search, where the metadata just consists of the selected metadata element or elements, in this case: Tags.

Metadata elements	Provider
CV	
DDC class labels	Amazon
Subject headings	Amazon
Geographic names	Amazon
Category labels	Amazon
LCSH terms	BL, LoC
Tags	
Tags	LT

Table 1: Metadata elements used in test collections and their respective providers.

Metadata elements
Individual
CV
Tags
Unique tags
Combined
Tags + CV
Unique tags + CV

Table 2: Test collections with different metadata elements.

In order to test the popularity effect of tags, we created an additional test collection Unique tags. This collection contains the same tags as in the Tags collection, but each tag is included only a single time in a book record.

To compare the complementarity effect of tags and CV terms, we also created two collections that are combinations of both sources: Tags + CV and Unique tags + CV. Table 2 contains an overview of the test collections that were used in this study.

3.2 Book Search Requests & Relevance Judgments

The book search requests that were used to search the collections were collected from LT discussion forums (Koolen, Kamps, & Kazai, 2012). Example requests include asking for (1) suggestions on books about a certain topic or from a particular genre; (2) ideas on books where the user can only remember plot details, but not the necessary metadata; and (3) recommendations based on individual preferences. Frequently, the requesters add books they have already read to their information need description. Figure 1 shows an example book request[5].

Book search requests taken from the LT forums consist of a title and a narrative, which were used in combination for searching in our experiments. Both can be considered realistic expressions of the information need. Section 5 describes these search requests in more detail.

The book suggestions posted as replies to the LT forum requests are regarded as relevant books for a request (Koolen et al., 2012). A graded relevance scale was used, based on additional criteria, such as whether the book had been added to the catalog of the requester or suggester(s) (Koolen, Kazai, Preminger, & Doucet, 2013). From the 2014 edition of the SBS workshop, we used 340 randomly selected topics for training and 334 topics for testing purposes, ensuring they were filtered to not include any of the ~800,000 book records that were filtered out.

3.3 Retrieval Setup & Evaluation

For our retrieval experiments, we used the Indri 5.4 toolkit[6] and its language modeling implementation with Jelinek-Mercer smoothing, which was shown to perform better on longer queries, such as the LT forum requests (Zhai & Lafferty, 2004).

To optimize the search system's performance for any of the five test collections, we used 340 randomly selected search requests and their relevant books for training to determine the system's parameter settings for (i) the degree of smoothing, as represented by the λ parameter, which controls the influence of the collection language model (varied in increments of 0.1, from 0.0 to 1.0); (ii) stopword filtering (none or using the SMART stopword list); and (3) stemming (none or using the Krovetz stemmer)[7]. These optimal settings were then used on the 334 test book search requests to produce the results presented in the remainder of this paper.

[5] Topic 99309, available at `http://www.librarything.com/topic/99309`, last accessed September 11, 2016.

[6] Available at `http://sourceforge.net/projects/lemur/files/lemur/indri-5.4/`

[7] This resulted in 44 different possible combinations of these three parameters, and $5 \times 44 = 220$ training runs in total. Readers interested in these optimal parameter settings are referred to `http://toinebogers.com/?page_id=738` for a complete overview.

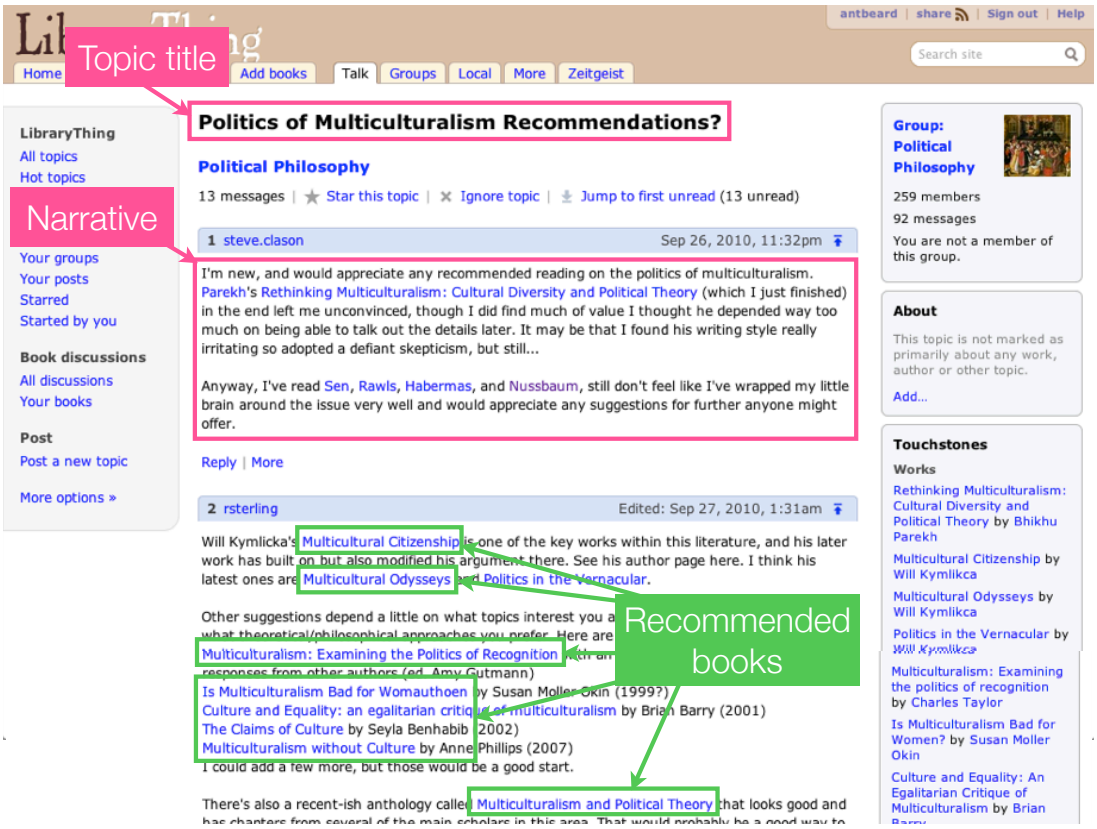

Figure 1: Book search request from the LibraryThing forums (re-used from (Bogers & Petras, 2015)).

We use NDCG@10 (Normalized Discounted Cumulated Gain cut off at rank 10, see Järvelin and Kekäläinen (2002)) to measure the search performance. This measure has also been used in previous SBS editions, making our work comparable. The single-figure NDCG@10 metric rewards result lists where highly relevant books are ranked higher.

We employ statistical significance testing when comparing the performance of different retrieval runs and use an α of 0.05. When comparing the performance of two different retrieval runs, we use two-tailed paired t-tests and also report the effect size (ES) and the 95% confidence interval (CI) as recommended by Sakai (2014).

4 A Comparative Analysis of Tags and Controlled Vocabularies

4.1 Main Results

Question 1: Is there a difference in performance between CV and Tags in retrieval?
Answer: Tags perform significantly better than CV . The combination of both sources in Tags + CV results in even better performance, but not significantly so.

Table 3 shows the main results of our five runs with Figure 2 representing the same information graphically. There is a statistically significant difference between the five runs according to a repeated-measures ANOVA with a Greenhouse-Geisser ($F(2.529, 842.144) = 4.650$, $p < .01$). We can see that Tags provide significantly better retrieval performance over our 334 requests compared to CV according to a two-tailed paired t-test ($t(333) = 2.171$, $p < .05$, ES = 0.118, 95% CI [0.0160, 0.0325]). However, combining the two in Tags + CV results in even better performance, which suggests they are complementary to a degree. While this combination also significantly outperforms the original CV collection ($t(333) = 2.874$, $p < .05$, ES = 0.157, 95% CI [0.0069, 0.0368]), Tags + CV does not perform significantly better than the Tags collection ($t(333) = 1.194$, $p = .253$, ES = 0.066, 95% CI [-0.0031, 0.1263]).

Metadata elements	NDCG@10
CV	0.0348
Tags	0.0519
Unique tags	0.0524
Tags + CV	0.0566
Unique tags + CV	**0.0583**

Table 3: Results for the different test collections using NDCG@10 as evaluation metric. The best-performing run is marked in bold font.

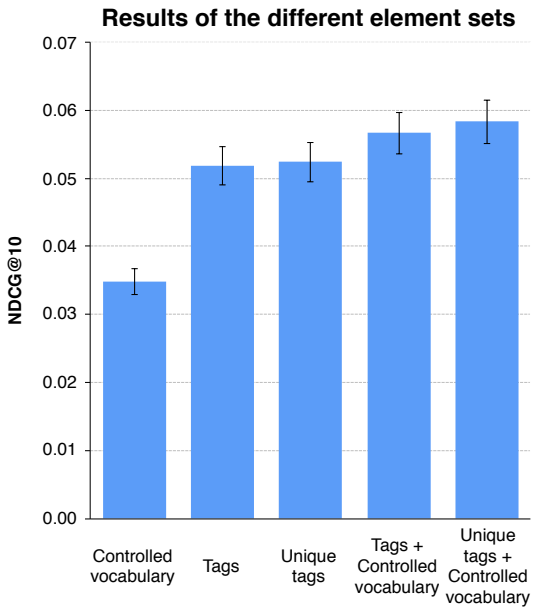

Figure 2: Results for the different test collections using NDCG@10 as evaluation metric. Bars indicate average NDCG@10 scores over all 334 topics, with error bars in black.

4.2 Popularity Effect

Question 2: Do Tags outperform CV because of the so-called popularity effect?
Answer: No, the popularity effect does not seem to be the reason for this difference. The Unique tags test collection (without tag frequency information) performs even better than Tags (albeit not significantly so).

One possible explanation for the difference between Tags and CV is the so-called popularity effect in tagging systems as first described by Noll and Meinel (2007): popular books received more (and more of the same) tags than unpopular books, whereas CV terms are more evenly distributed across books. In our previous study (Bogers & Petras, 2015), we also put forward this phenomenon as a possible explanation for the performance difference, similar to Koolen (2014).

One component of this popularity effect is books receiving more unique tags than CV terms, which requires an examination of the type and token statistics of our collections. Table 4 shows type and token counts for the different collections, both as total counts and averages per document. It shows that books do *not* receive more unique tags (= types) per document than CV terms. In fact, the average number of types assigned per document for CV is nearly three times higher at 36.52 compared to 13.08 for Tags. The high number of types for CV can be explained by the aggregation of several controlled vocabulary metadata fields in one test collection, as explained by table 1.

Metadata elements	#types	#tokens	avg. types/doc	avg. tokens/doc
CV	2,208,694	109,793,695	36.52	53.3
Tags	2,272,393	246,313,480	13.08	119.5
Unique tags	2,272,393	47,253,002	13.08	22.9
Tags + CV	2,353,659	354,046,417	29.43	171.8
Unique tags + CV	2,353,659	154,985,939	29.43	75.2

Table 4: Type and token statistics for the five different metadata element sets.

The other aspect of the popularity effect is that the books receive more of the same tags, i.e., that tag frequency

plays an import role in the performance difference. Table 4 does show that the average number of tokens assigned to a document is more than twice as high for Tags at 119.5 than it is for CV at 53.5.

To further examine this, we created our Unique tags collection, removing tag frequency information from the Tags collection and only retaining single occurrences of each tag. If tag frequency is indeed the deciding factor, then Tags should perform better than Unique tags . However, the opposite is true: Unique tags achieves better performance than Tags, even though the difference is not statistically significant according to a two-sided paired-samples t-test ($t(333) = 0.139$, $p = .890$, = 0.007, 95% CI [-0.0070, 0.0080]). Moreover, Unique tags shows an even bigger, statistically significant performance increase over CV than Tags did ($t(333) = 2.135$, $p < .05$ (0.033), = 0.117, 95% CI [0.0014, 0.0338]). This strongly suggests that it is the quality of the tags themselves that makes the difference with CV instead of a popularity effect. On average, CV simply do not appear to match the user's vocabulary as well.

4.3 Complementarity

Question 3: Do Tags, Unique tags, and CV complement or cancel each other out in terms of retrieval performance?
Answer: Tags, Unique tags, and CV complement each other: they are successful on different sets of requests.

The best-performing of our five runs is the Unique tags + CV collection. The success of this combination suggests that the two individual representations Unique tags and CV provide the best complementary performance. This is also confirmed by the per-request difference plots in Figure 3.

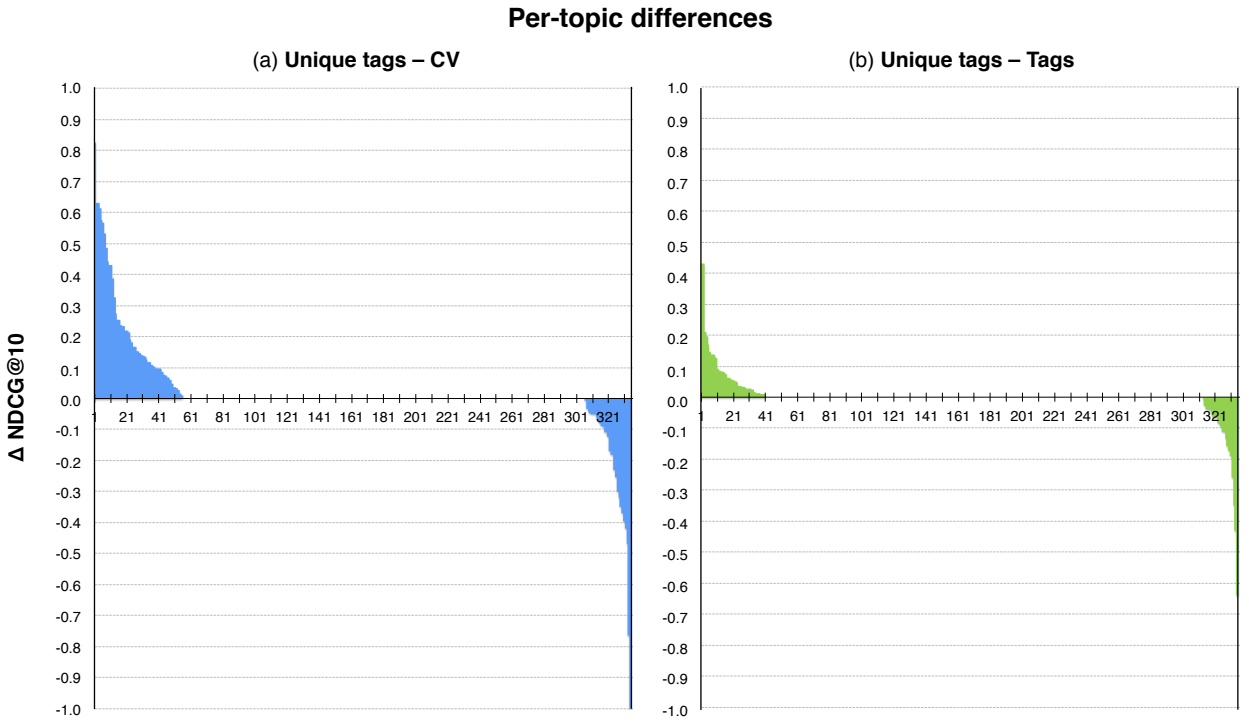

Figure 3: Differences in retrieval performance ordered by per-request difference between (a) the Unique tags and CV, and (b) the Unique tags and Tags collections. Bars above the horizontal axis represent requests where Unique tags perform better, bars below the horizontal axis represent requests where the other collections perform better.

It shows the per-request differences in NDCG@10 between (a) a search using Unique tags and CV; and (b) a search using Unique tags and Tags.

Figure 3a shows how many of the requests were better served by Unique tags (bars above the horizontal line) and how many requests were better served by retrieving using the CV (bars below the horizontal line). As the area above the horizontal axis is larger than the area below it, this again demonstrates that Unique tags show a

small advantage over CV. It also shows that there are different types of search requests: for most requests, searching either Unique tags or CV makes no difference, but for certain requests one of the two test collections outperforms the other. The pattern in Figure 3a confirms that Unique tags and CV indeed offer complementary performance.

We see an identical pattern for the difference between Unique tags and Tags in Figure 3b. This suggests that while, on average, tag frequency information hurts retrieval performance slightly more than it helps, there are also several requests where tag frequency information actually improves the ranking enough so that Tags outperforms Unique tags .

Despite their complementarity, the majority of requests shown in Figures 3a show no performance difference between Unique tags [8] and CV. Most of these requests actually failed to retrieve any relevant documents at all: 247 out of 334 test topics (or 74.0%) fail completely.

This leads us to two very interesting follow-up questions: (1) what is it in these Unique tags and CV collections that helps successfully retrieve relevant documents, and (2) what makes the overwhelming majority book search requests so difficult that both metadata elements fail completely at retrieving relevant documents? We attempt to answer these two questions in Sections 5 and 6 respectively.

5 Analysis of Book Search Requests

In the previous section we learned that despite overall performance differences, Unique tags and CV terms offer complementary performance. In this section, we take a closer look at 87 requests (or 26.0%) that succeeded in at least one relevant document being retrieved in the top 10 results by Unique tags or CV. What makes certain representations better at satisfying some types of book search requests than others?

5.1 Relevance Aspects in Book Search Requests

Question 4: What types of book requests (in terms of what makes them relevant to users) are best served by the Unique tags and CV test collections?
Answer: CV show a tendency to work best for requests that touch upon aspects of engagement, whereas requests that focus on content-based, familiarity, known-item, or socio-cultural aspects are best served by Unique tags.

One way of categorizing book search requests is by the relevance aspects that are expressed in it: what aspects make a book relevant to the original poster? While some LT users are trying to re-find a book from their childhood with only vague plot points and memories of characters to go on, others express a desire for books that match a specific mood or provide a certain reading experience.

To analyze the difference between Unique tags and CV in terms of such relevance aspects expressed in the requests, we use the relevance aspects identified and annotated by Koolen et al. (2015) and inspired by Reuter (2007). They annotated a large set of SBS book requests (which include our 334 test requests as a subset) with one or more of a set of eight relevance aspects[9]. Table 5 contains brief descriptions of these eight relevance aspects.

It shows that among the successful requests, Content is the most common relevance aspect in 79.3% of all 87 topics, followed by Familiarity and Metadata.

If we compare the search requests where one of the test collections outperforms the other by a margin of at least 120%, then we see a clear difference in the distribution of aspects. Apart from Engagement, which is best served by CV representations, all other aspects are best satisfied with Unique tags .

Figure 4 shows how well the two test collections perform on requests of different types as measured by NDCG@10, and it shows largely the same pattern as Table 5. While all aspects except Engagement perform better when using Unique tags , the difference between Unique tags and CV is only statistically significant for Familiarity according to a two-tailed paired-samples t-test ($t(35) = 2.268$, $p < .05$, ES = 0.377, 95% CI [0.0119, 0.2147]) and for Content ($t(62) = 3.489$, $p < .005$, ES = 0.440, 95% CI [0.0489, 0.1800]).

When inspecting the relevant documents, it is easy to see why these patterns occur. A good example for the Content aspect is topic #63529 ("*I just finished and enjoyed Climb the Wind by Pamela Sargent. Can anyone recommend other science fiction and or alternate history about Native Americans?*"). Several of the relevant retrieved documents are

[8] In the remainder of this paper we will use Unique tags as our collection representing tags, because they provide the best individual performance.

[9] Available at `http://social-book-search.humanities.uva.nl/#/data/suggestion`, last visited September 16, 2016.

Relevance aspect	Description	Requests overall (N = 87)	UniqueTags > CV (N = 53)	CV > UniqueTags (N = 27)
Accessibility	Language, length, or level of difficulty of a book	9.2%	7.5%	11.1%
Content	Topic, plot, genre, style, or comprehensiveness	79.3%	83.0%	70.4%
Engagement	Fit a certain mood or interest, are considered high quality, or provide a certain reading experience	25.3%	22.6%	33.3%
Familiarity	Similar to known books or related to a previous experience	47.1%	49.1%	37.0%
Known-item	The user is trying to identify a known book, but cannot remember the metadata that would locate it	12.6%	17.0%	7.4%
Metadata	With a certain title or by a certain author or publisher, in a particular format, or certain year	23.0%	24.5%	14.8%
Novelty	Unusual or quirky, or containing novel content	3.4%	3.8%	0%
Socio-cultural	Related to the user's socio-cultural background or values; popular or obscure	13.8%	15.1%	7.4%

Table 5: Distribution of the relevance aspects over all 87 successful book requests (column 1), the requests where Unique tags outperform CV terms by 120% or more (column 2), and the requests where CV terms outperform Unique tags by 120% or more (column 3). More than one aspect can apply to a single book request, so numbers to not add up to 100%.

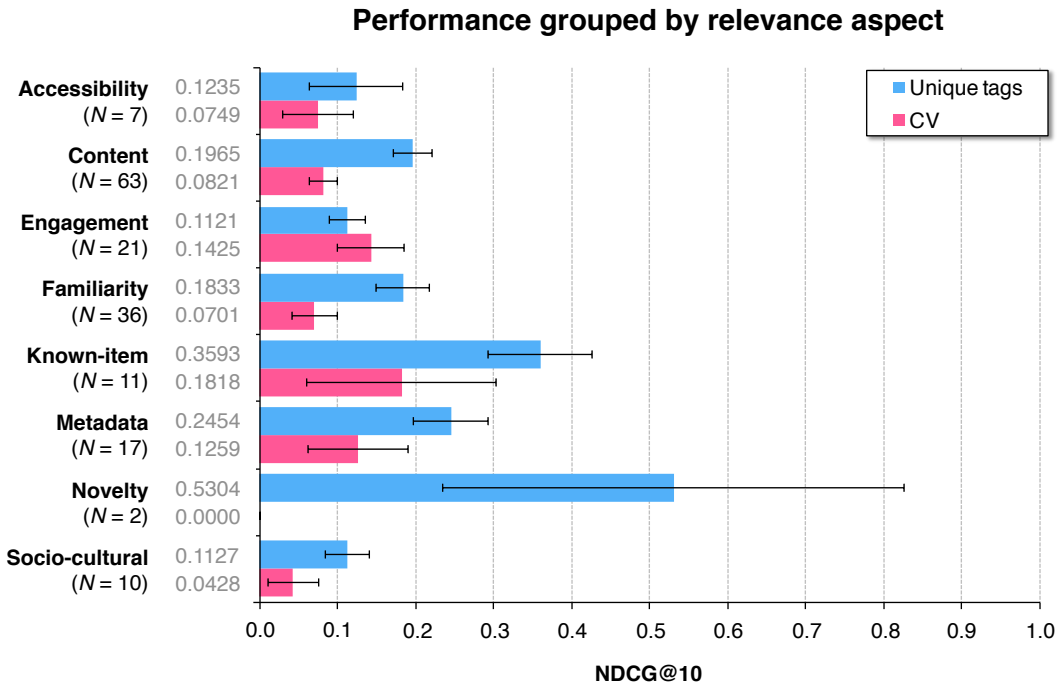

Figure 4: Results for the Unique tags and CV test collections, grouped by the eight relevance aspects expressed in the 87 successful book search requests. Average NDCG@10 scores over all requests expressing a particular relevance aspect are shown in grey and as horizontal bars, with error bars in black.

only indexed with Science fiction in CV , but not with terms like alternate history and native americans, which *are* present in Unique tags and greatly improve their chances of being ranked at the top of the results list.

Known-item requests are by their very nature difficult to fulfill using CV as the user typically only remembers some vague plot elements and characters, which are not the most important topics that subject headings tend to cover. For example, in request #73796 ("*I read this book 5 to 10 years ago. It was like Francine Rivers, but doesn't seem to match any of her titles that I can find. It started with 3 older men of a small church searching for a new pastor and hiring a young man who seemed promising. The new pastor had great success but as the church grew into a mega church with building projects, etc, he strayed away from the Word.*"), the tags church growth, pastor, mega churches are what rank the relevant document near the top. The more generic CV terms Church buildings and Clergy are not enough to provide effective retrieval.

Other aspects like Socio-cultural and Novelty are also more likely to be present as tags than as CV, resulting in their improved performance with Unique tags. Given typical indexing guidelines, one could perhaps expect that Accessibility and Content aspects would be covered well by CV . Figure 4 shows that this is not the case. Perhaps surprisingly, it is actually the Engagement topics that are better served on average by CV than by Unique tags. This difference, however, is not significant ($t(20) = 0.767$, $p = .452$, ES = 0.167, 95% CI [-0.1132, 0.0524]), and an inspection of the Engagement requests revealed no CV or Unique tags terms related to reading engagement, suggesting that the difference is coincidental.

5.2 Book Type: Fiction vs. Non-fiction

Question 5: What types of book requests (in terms of fiction or non-fiction books that are requested) are best served by Unique tags or CV ?
Answer: Unique tags work much better than CV for fiction book requests. CV show a tendency to work better for non-fiction book requests, but the difference is not significant.

Search requests typically ask for one of two types of books: fiction or non-fiction. For our previous study (Bogers & Petras, 2015), we annotated all 334 requests as for works of fiction or non-fiction. Fiction and non-fiction requests are unevenly distributed in our test set: the majority of the requests (75.3%) were for works of fiction.

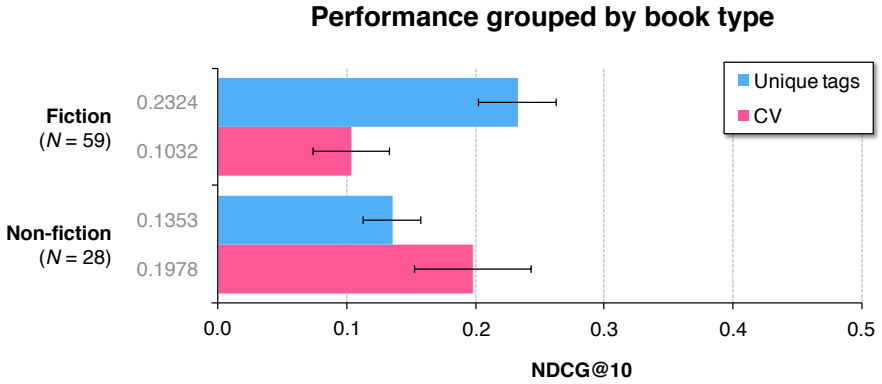

Figure 5: Results for the Unique tags and CV test collection, grouped by type of book(s) requested (fiction or non-fiction). Average NDCG@10 scores over all requests for a particular book type are shown in grey and as horizontal bars, with error bars in black.

An analysis of the performance of Unique tags and CV with respect to the nature of the book(s) being requested (see Figure 5) shows that Unique tags are significantly better in serving requests for fiction books than CV ($t(58) = 3.571$, $p < .005$, ES = 0.465, 95% CI [0.0568, 0.2016]). While CV is better than Unique tags in serving non-fiction requests, this difference is not statistically significant ($t(27) = 1{,}194$, $p = .243$, ES = 0.226, 95% CI [-0.1699, 0.0449]).

What could be the reason for the large difference between Unique tags and CV for fiction requests? To explain this, we can consult the distribution of relevance aspects by the type of book(s) requested in Figure 6. Aspects that are more commonly expressed in requests for fiction are exactly those aspects that Unique tags tend to be better at solving. For example, the Known-item aspect occur more often in fiction requests: 41.0% of all fiction requests

cover this aspect versus 30.8% of non-fiction requests. As we saw in Figure 4 in the previous section, these are more likely to be solved by Unique tags, which explains (part of) the better performance on fiction requests.

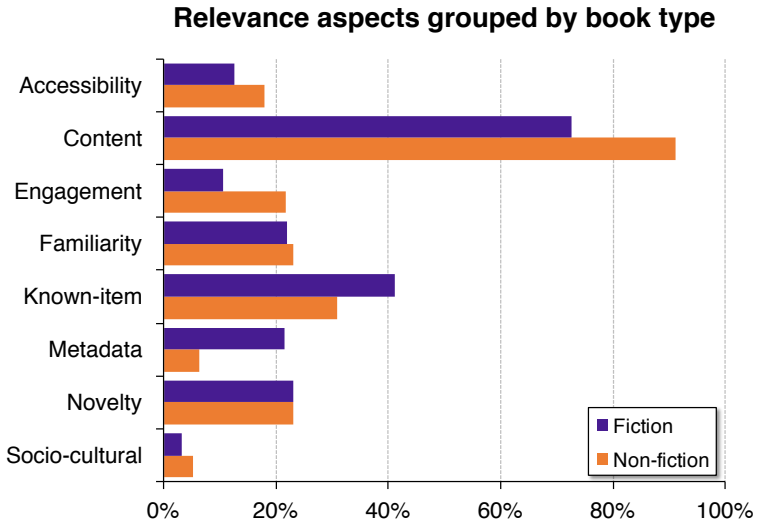

Figure 6: Distribution of relevance aspects by the type of book(s) requested (fiction (N = 256) vs. non-fiction (N = 78). Horizontal bars represent the percentage of all request of a particular book types that express a specific aspect. For example, 41.0% of all 256 fiction requests express a Known-item aspect.

Metadata is another aspect that is more common in fiction requests at 21.5% versus 6.4% of non-fiction requests. Normally, requests for books from a certain author or from a specific publication year would by solved by core bibliographic metadata. However, in this pure comparison between Unique tags and CV the former performs better, because in free tagging schemes these metadata elements will inevitably be added as tags by some users, whereas CV will not describe them. Claiming this as an advantage for tags is unfair, however, as any normal digital library search engine would index such core bibliographic metadata regardless.

Perhaps counter-intuitively, aspects such as Content, Engagement, and Accessibility are more common in non-fiction requests. Engagement was the one aspect that was better addressed by CV (albeit not significantly), which could help explain the better performance of CV on non-fiction requests. Requests that express a Content aspect are also understandably tied to non-fiction requests, which commonly include elements like the topic and the degree of comprehensiveness. All of this suggests that the relative benefits of Unique tags and CV are strongly dependent on the types of book requests made and on the aspects of books relevant to the requester.

6 Failure Analysis

Despite the complementarity of Unique tags and CV , the majority of book requests fail: 74.0% of all 334 book search requests fail to retrieve any relevant books. Combining the two test collections in Unique tags + CV produces non-zero results for 7 more requests, due to improved ranking, but for 240 requests even Unique tags + CV fails to retrieve relevant documents in the top 10 results. In this section, we perform a failure analysis of these 240 requests: What kind of requests fail most frequently? What is the reason for this? And is this a problem that tagging and controlled vocabularies could ever be expected to solve?

6.1 Sparsity, Recall Base, and Example Books

One common cause of poor performance in retrieval and recommendation systems is data sparsity: many algorithms breakdown when not provided with enough data. Sparsity could affect book retrieval in two different ways: (1) book search requests could be too short and thereby provide inadequate information to locate relevant documents, and (2) book metadata could be too short, making it difficult to match them against rich book search requests.

Question 6: Do failed book search requests fail because of data sparsity, a lower recall base, or a lack of examples?
Answer: Sparsity does not appear to be a reason for retrieval failure and neither is the size of the recall base. The number of examples provided by the requester does have a significant positive influence on performance.

Table 6 shows the average length of search requests and relevant documents for successful and failed requests. This suggests that the length of search requests is unlikely to be the underlying cause. Not only is there no significant difference between the two groups according to an independent-samples t-test ($t(332) = 0.907$, $p = .365$, 95% CI [9.915, 10.933]), but the difference actually goes the other way as failed requests are longer on average than successful ones. Document length does not appear to be the reason either: the relevant documents for failed search requests are longer on average and statistically significantly so ($t(3889) = 6.257$, $p < .001$, 95% CI [-5.580, 0.892]). Instead of sparsity, a possible cause could be that the retrieval algorithm is unable to distinguish well enough between important and unimportant terms and that request and document length exacerbate this problem.

	Avg. book search request length (in words)	Avg. relevant document length (in words)	Avg. no. of relevant documents	Avg. no. of example books provided
Successful ($N = 94$)	86.7	73.9	13.3	1.63
Failed ($N = 240$)	96.6	79.5	11.0	0.54
Overall ($N = 334$)	93.8	77.7	11.7	0.84

Table 6: Breakdown of book search requests by request length, length of the relevant documents, size of the recall base, and the number of examples provided by the original requester.

Another possible reason for retrieval success is the recall base: more relevant documents means it is relatively easier to return relevant documents in the results list. Again the numbers in Table 6 do not bear this out: the two groups have an average difference of only 2.3 documents, which is not statistically significant ($t(332) = 1.269$, $p = .205$, 95% CI [-2.301, 1.812]).

Finally, the number of examples provided by the original requester may influence performance: with more relevant examples, the retrieval engine could potentially provide better results. This explanation appears to have some merit. There is a significant difference in the number of examples provided for successful and failed requests ($t(332) = 4.638$, $p < .001$, 95% CI [-1.098, 0.237]), as shown in Table 6. There is also a weak positive correlation $r = 0.175$ ($p < .005$) between NDCG@10 score for Unique tags + CV and the number of provided examples.

6.2 Relevance Aspects and Book Types

Question 7: Do book search requests fail because of their relevance aspects?
Answer: No. The relevance aspects are distributed equally for successful and failed requests. Only Accessibility and Metadata related search requests seem to fail more often.

Figure 4 in Section 5 showed that some of the performance differences between search requests could be explained by the different relevance aspects that are expressed in them. For example, Known-item and Metadata requests achieve higher NDCG@10 scores than Accessibility and Socio-cultural requests. A possible explanation for the failed requests could be that they contain proportionately more of the difficult aspects. Table 7 and Figure 6 show the distribution of relevance aspects over the successful and failed requests.

Relevance aspect	Successful (N = 94)	Failed (N = 240)
Accessibility	9.6%	15.4%
Content	79.8%	75.8%
Engagement	14.9%	12.5%
Familiarity	24.5%	21.3%
Known-item	45.7%	35.8%
Metadata	13.8%	19.6%
Novelty	24.5%	22.5%
Socio-cultural	4.3%	3.3%

Table 7: Tabular distribution of the relevance aspects over all 94 successful and 240 failed requests.

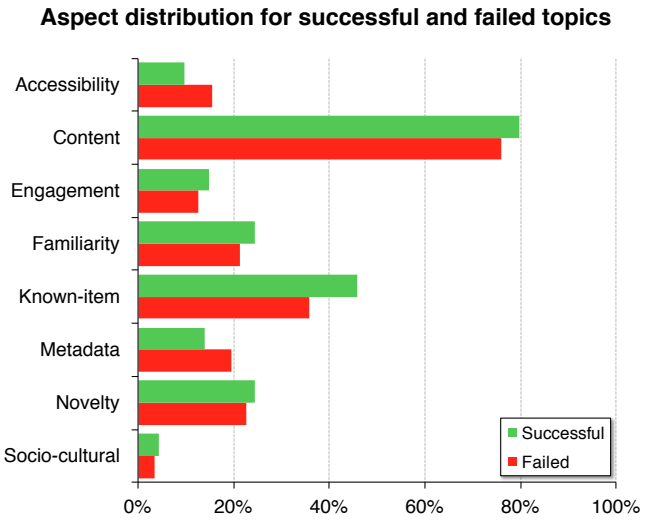

Figure 7: Visual distribution of the relevance aspects over all 94 successful and 240 failed requests.

The two distributions for successful and failed requests do not differ greatly from one another. Accessibility and Metadata occur in greater proportion among failed requests, suggesting that these are among the harder types to solve. While the 13 Metadata requests that successfully retrieve relevant documents tend to achieve good results, the majority of Metadata requests (N = 47) actually fail to retrieve any relevant documents.

The same holds for the Known-item requests: while 42 requests are successful, more than double that number of requests—86 in total—fail to rank any relevant documents in the top 10 results. This suggests that even for Unique tags , these request types are far from a solved problem.

Question 7: Does the type of book that is being requested (fiction vs. non-fiction) have an influence on whether requests succeed or fail?
Answer: Requests for works of fiction result significantly more often in failed book search requests.

Finally, another way of categorizing our book search requests is by the type of book requested: fiction or non-fiction. Of the 240 failed requests, 193 (or 80.4%) were for fiction, whereas only 63 out of 94 successful requests (or 67.0%) were for fiction. This difference is statistically significant according to a Chi-square test ($\chi^2(1) = 6.771$, $p < .01$) and shows that fiction requests tend to be harder to solve.

7 Discussion & Conclusions

In this paper we have presented a large-scale empirical comparison and in-depth analysis of the performance of controlled vocabulary vs. social tagging metadata for book search. Using a large collection of book records and book search requests that go beyond simple bibliographic queries, we showed that tags offer a richer vocabulary for answering complex book search requests than CV terms in general, but that the performance of the two representations is complementary. We also provided compelling evidence against the often suggested popularity effect of tags as the reason for their superior performance. Instead, it is the quality of matches between user-provided tags and search requests that results in better performance. A detailed request analysis showed that the relative performance of tags and CV appears to be dependent on the type of request, both in terms of the relevance aspects expressed in them and the types of works being requested. These factors appear to be more predictive of which representation performs best.

Finally, a comparative analysis of successful and failed search requests showed that addressing complex search requests using book search engines is far from a solved problem as most requests fail to retrieve relevant documents. Especially fiction book search requests are hard to fulfill, probably because requesters are asking for more aspects (engagement, familiarity, plot details) that are less covered by tags and possibly not at all by CV. We posit from

these results that existing indexing practices for books will have to change if complex book search requests are to have a chance of being met. Plot details could be added by harvesting the full-text of books for keywords. Indeed, Amazon already adds character and place names from its books to some of its controlled metadata. A detailed genre classification, a reading speed or engagement level estimate are highly subjective data points, which is why controlled vocabularies have avoided to use them for indexing. However, this type of information is often exactly what searchers are looking for. A critical look at existing subject indexing guidelines is required if we want book search engines to solve all book-related information needs and not just the simple ones.

For future work, we are considering developing a predictive model of query difficulty that takes all of the relevant factors into account. Another interesting avenue of research would be to examine the completeness of the current set of relevance judgments: currently, these are restricted by the suggestions made by other users, but an inspection of the data suggests these may be incomplete.

References

Bartley, P. (2009). Book Tagging on LibraryThing: How, Why, and What are in the Tags? *Proceedings of the American Society for Information Science and Technology*, *46*(1), 1–22.

Beckers, T., Fuhr, N., Pharo, N., Nordlie, R., & Fachry, K. N. (2010). Overview and Results of the INEX 2009 Interactive Track. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, & I. Frommholz (Eds.), *Research and Advanced Technology for Digital Libraries* (Vol. 6273, p. 409-412). Springer.

Bogers, T., & Petras, V. (2015). Tagging vs. Controlled Vocabulary: Which is More Helpful for Book Search? In *Proceedings of iConference 2015.*

Buchanan, G., & McKay, D. (2011). In the Bookshop: Examining Popular Search Strategies. In *JCDL '11: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 269–278). New York, NY, USA: ACM.

Buckley, C. (2009). Why Current IR Engines Fail. *Information Retrieval*, *12*(6), 652–665.

Carmel, D., & Yom-Tov, E. (2010). Estimating the Query Difficulty for Information Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *2*(1), 1–89.

Cleverdon, C. W., & Mills, J. (1963). The Testing of Index Language Devices. In *Aslib proceedings* (Vol. 15, pp. 106–130).

Dextre Clarke, S. (2008). The Last 50 Years of Knowledge Organization: A Journey through my Personal Archives. *Journal of Information Science*, *34*(4), 427-437.

Dextre Clarke, S., & Vernau, J. (2016). The Thesaurus Debate Continues. *Knowledge Organization*, *43*(3), 135–137.

Gross, T., & Taylor, A. G. (2005). What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results. *College & Research Libraries*, *66*(3), 212-30.

Heymann, P., & Garcia-Molina, H. (2009). Contrasting Controlled Vocabulary and Tagging: Do Experts Choose the Right Names to Label the Wrong Things? In *WSDM '09: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (Late Breaking Results Session)* (pp. 1–4). Stanford InfoLab.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446.

Kim, J. Y., Feild, H., & Cartright, M. (2012). Understanding Book Search Behavior on the Web. In *CIKM '12: Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 744–753). New York, NY, USA: ACM.

Koolen, M. (2014). "User Reviews in the Search Index? That'll Never Work!". In *ECIR '14: Proceedings of the 36st European Conference on Information Retrieval* (pp. 323–334).

Koolen, M., Bogers, T., Gäde, M., Hall, M., Hendrickx, I., Huurdeman, H., … Walsh, D. (2016). Overview of the CLEF 2016 Social Book Search Lab. In *CLEF 2016: Proceedings of the 7th International Conference of the CLEF Association* (Vol. LNCS 9822, p. 351-370).

Koolen, M., Bogers, T., Van den Bosch, A., & Kamps, J. (2015). Looking for Books in Social Media: An Analysis of Complex Search Requests. In A. Hanbury, G. Kazai, A. Rauber, & N. Fuhr (Eds.), *ECIR 2015: Proceedings of the 37th European Conference on IR Research* (pp. 184–196). Springer.

Koolen, M., Kamps, J., & Kazai, G. (2012). Social Book Search: Comparing Topical Relevance Judgements and Book Suggestions for Evaluation. In *CIKM '12: Proceedings of the 21st International Conference on Information and Knowledge Management* (pp. 185–194).

Koolen, M., Kazai, G., Preminger, M., & Doucet, A. (2013). Overview of the INEX 2013 Social Book Search Track. In P. Forner, J. Karlgren, & C. Womser-Hacker (Eds.), *CLEF 2013: Proceedings of the Fourth International Conference of the Cross-Language Evaluation Forum* (pp. 1–26).

Lawson, K. G. (2009). Mining Social Tagging Data for Enhanced Subject Access for Readers and Researchers. *The Journal of Academic Librarianship*, *35*(6), 574 - 582.

LoC Cataloging Policy and Support Office. (2016). Assigning and Constructing Subject Headings H 180. In *Library of Congress Subject Heading Manual.* Retrieved from https://www.loc.gov/aba/publications/FreeSHM/H0180.pdf

Lu, C., Park, J.-R., & Hu, X. (2010). 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.

Magdy, W., & Darwish, K. (2008). Book Search: Indexing the Valuable Parts. In *BooksOnline '08: Proceedings of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories* (pp. 53–56). New York, NY, USA: ACM.

Mikkonen, A., & Vakkari, P. (2016). Readers' Interest Criteria in Fiction Book Search in Library Catalogs. *Journal of Documentation*, *72*(4), 696-715.

Noll, M. G., & Meinel, C. (2007). Authors vs. Readers: A Comparative Study of Document Metadata and Content in the Www. In *DocEng '07: Proceedings of the 2007 ACM Symposium on Document Engineering* (pp. 177–186). New York, NY, USA: ACM.

Phillips, H. (2010). The Great Library of Alexandria? *Library Philosophy and Practice*, *August*. Retrieved from http://unllib.unl.edu/LPP/phillips.pdf

Reuter, K. (2007). Assessing Aesthetic Relevance: Children's Book Selection in a Digital Library. *Journal of the American Society for Information Science and Technology*, *58*(12), 1745–1763.

Rowley, J. E. (1994). The Controlled versus Natural Indexing Languages Debate Revisited: A Perspective on Information Retrieval Practice and Research. *Journal of Information Science*, *20*(2), 108-19.

Saarinen, K., & Vakkari, P. (2013). A Sign of a Good Book: Readers' Methods of Accessing Fiction in the Public Library. *Journal of Documentation*, *69*(5), 736-754.

Sakai, T. (2014). Statistical Reform in Information Retrieval? *SIGIR Forum*, *48*(1), 3–12.

Slone, D. J. (2000). Encounters with the OPAC: On-line Searching in Public Libraries. *Journal of the American Society for Information Science*, *51*(8), 757–773.

Smith, T. (2007). Cataloging and You: Measuring the Efficacy of a Folksonomy for Subject Analysis. In J. Lussky (Ed.), *Proceedings 18th Workshop of the ASIST Special Interest Group in Classification Research.* Retrieved from http://dlist.sir.arizona.edu/2061/

Voorbij, H. (2012). The Value of LibraryThing Tags for Academic Libraries. *Online Information Review*, *36*(2), 196-217.

Willis, C., & Efron, M. (2013). Finding Information in Books: Characteristics of Full-text Searches in a Collection of 10 Million Books. In *ASIST '13: Proceedings of the 76th ASIS&T Annual Meeting* (pp. 84:1–84:10). Silver Springs, MD, USA: American Society for Information Science.

Zhai, C., & Lafferty, J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, *22*(2), 179–214.

Table of Figures

Table of Tables