Can Tagging Organize Human Knowledge?

InfoLab Technical Report, Last Updated November 15, 2008

Paul Heymann and Hector Garcia-Molina
Department of Computer Science
Stanford University, Stanford, CA, USA
{heymann, hector}@cs.stanford.edu

ABSTRACT
The fundamental premise of tagging systems is that regular users can successfully organize large collections for browsing and other tasks through the use of uncontrolled vocabularies. Until now, that premise has remained relatively unexamined. In this work, using library data as a guide, we look at whether tags do a good job of organizing large sets of objects. We find that tags are predominantly objective and content-based, that synonymy is not a huge problem, and that the groups of objects are well distributed in size. We find that tags often have equivalent or contained library terms. Using these equivalent and contained terms, we examine the recall of a tag on the objects it could annotate. We find that tags have high recall on popular works, but lower recall on less popular works. Lastly, we find that tags have some desirable traits for information integration, like similar top tags across systems and similar annotations of the same object across systems.

1. INTRODUCTION
Over the past five years, collaborative tagging systems (like Flickr and delicious) have become one of the dominant means of organizing knowledge and data on the web. These systems allow any user to annotate any object (like a photo or bookmark) with a “tag” (i.e., a keyword) from an uncontrolled vocabulary. Over time, the most popular tags mature into a vocabulary specially adapted to the contents of the system. Furthermore, due to the focus on user (rather than expert) annotation, collaborative tagging systems are highly scalable, without the system owner needing to expend resources on taxonomists and catalogers for the data.

These features make collaborative tagging particularly well suited for the web. A popular website often has many users, unknown future objects, and few resources dedicated to data organization. In this context (and given ease of implementation), it is no surprise that collaborative tagging systems have become as popular as they have.

No one has analyzed the fundamental premise of these systems until now. Are tags really organizing data as well as a taxonomist or cataloger might? Are they really adapted to the objects in the system? Previous work has shown, for example, that some tags do become much more popular than others. However, no one has really asked: are these tags any good for organizing the data they are meant to organize? The answer to this question impacts how some of the most important and most popular sites on the web with millions or perhaps even billions of objects in their collaborative tagging systems organize their data, and whether that organization makes sense.

In this paper, we look at this question through the lens of libraries. When we say that we are curious how well tags “organize” data, what we mean is how well tags facilitate browsing over the data. In other words, if you did not know in advance what you were looking for, would tags help you find it? This is a very difficult problem to quantify. One cannot simply give users a description of an object to find, because then they already know what they are looking for!

Libraries have been working to solve this problem for decades. Tools like classifications (e.g., Dewey Decimal Classification and Library of Congress Classification) and subject headings (e.g., Library of Congress Subject Headings) are expressly designed to enable browsing. Library metadata is thus very close to a gold standard for organizing books to enable browsing. We wonder if tags can organize knowledge as well as library terms, but without the cost.

We look at data from two social cataloging sites, LibraryThing and Goodreads. Social cataloging sites are collaborative tagging systems where users tag books. How similar are the groups of books developed by these tagging systems over the course of a few years to the gold standard that librarians have built over decades?

In Section 2 we build a vocabulary to discuss tagging and library data. In Section 3, we break the question “how well do tags organize data?” into seven major features which we believe gold standard library records have, that tagging might emulate. In Section 4, we describe our dataset for investigating this question. In each subsection of Section 5, we address one of the seven features and determine if tagging systems have it. Throughout these experiments, we use a collection of methods ranging from association rules to Mechanical Turk workers to indexed searches. We find that tags do a remarkable job of emulating many of the features that libraries use to organize data, suggesting that tags are a high quality tool for data organization.

2. GENERAL PRELIMINARIES
A social tagging system consists of users \( u \in U \), annotations \( a \in A \), and objects \( o \in O \). In this paper we focus on social cataloging sites where the objects are books. More accurately, an object is a work, which represents one or more closely related books (e.g., the different editions of a book represent a work).

An object \( o \) can be annotated in two ways. Object \( o \) can be annotated by a user of the site, in which case we call the annotation a tag (written \( t_i \in T \)). For example, a user can tag a work with “interesting” or with “science
Table 1: Top 15 Tags (“Type” from Section 5.1).

<table>
<thead>
<tr>
<th>Tag</th>
<th>Type</th>
<th>Count</th>
<th>Tag</th>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-fiction</td>
<td>obj.</td>
<td>68974</td>
<td>biography</td>
<td>obj.</td>
<td>19409</td>
</tr>
<tr>
<td>fiction</td>
<td>obj.</td>
<td>48671</td>
<td>novel</td>
<td>obj.</td>
<td>18363</td>
</tr>
<tr>
<td>history</td>
<td>obj.</td>
<td>42447</td>
<td>tbr acrn.</td>
<td>obj.</td>
<td>18260</td>
</tr>
<tr>
<td>read</td>
<td>?</td>
<td>33271</td>
<td>20th century</td>
<td>obj.</td>
<td>17756</td>
</tr>
<tr>
<td>unread</td>
<td>?</td>
<td>31911</td>
<td>literature</td>
<td>obj.</td>
<td>17735</td>
</tr>
<tr>
<td>own</td>
<td>?</td>
<td>27377</td>
<td>wishlist</td>
<td>pers.</td>
<td>16508</td>
</tr>
<tr>
<td>reference</td>
<td>obj.</td>
<td>21953</td>
<td>hardcover</td>
<td>phys.</td>
<td>13987</td>
</tr>
<tr>
<td>paperback</td>
<td>phys.</td>
<td>20162</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Top 10 LC Classes.

<table>
<thead>
<tr>
<th>Description</th>
<th>LC</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language &amp; Literature</td>
<td>P</td>
<td>98410</td>
</tr>
<tr>
<td>Philosophy, Psychology, &amp; Religion</td>
<td>B</td>
<td>37399</td>
</tr>
<tr>
<td>American Literature</td>
<td>PS</td>
<td>30832</td>
</tr>
<tr>
<td>Fiction and juvenile belles lettres</td>
<td>PZ</td>
<td>27063</td>
</tr>
<tr>
<td>Individual American Authors</td>
<td>PS 700-3577</td>
<td>26296</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>H</td>
<td>24722</td>
</tr>
<tr>
<td>Technology</td>
<td>T</td>
<td>22312</td>
</tr>
<tr>
<td>Science &amp; Math</td>
<td>Q</td>
<td>21875</td>
</tr>
<tr>
<td>Juvenile belles lettres</td>
<td>PZ 5-91</td>
<td>21863</td>
</tr>
<tr>
<td>American Literature 1961-2000</td>
<td>PS 3550-3577</td>
<td>21302</td>
</tr>
</tbody>
</table>

Table 3: Top 10 Dewey Decimal Classes.

<table>
<thead>
<tr>
<th>Description</th>
<th>DDC</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature &amp; Rhetoric</td>
<td>8xx</td>
<td>63420</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>3xx</td>
<td>44663</td>
</tr>
<tr>
<td>Geography &amp; History</td>
<td>9xx</td>
<td>38325</td>
</tr>
<tr>
<td>The Arts</td>
<td>7xx</td>
<td>36300</td>
</tr>
<tr>
<td>Technology (Applied Sciences)</td>
<td>6xx</td>
<td>36230</td>
</tr>
<tr>
<td>American &amp; Canadian Literature</td>
<td>81x</td>
<td>33699</td>
</tr>
<tr>
<td>Religion</td>
<td>2xx</td>
<td>27124</td>
</tr>
<tr>
<td>American &amp; Canadian Fiction</td>
<td>813</td>
<td>25033</td>
</tr>
<tr>
<td>English &amp; Related Literature</td>
<td>82x</td>
<td>16670</td>
</tr>
<tr>
<td>Natural Sciences &amp; Mathematics</td>
<td>5xx</td>
<td>15960</td>
</tr>
</tbody>
</table>

2.1 Library Terms

We look at three types of library terms: classifications, subject headings, and the contents of MARC 008.

A classification is a set of annotations arranged as a tree, where each annotation may contain one or more other annotations. An object is only allowed to have one position in a classification. This means that an object is associated with one most specific annotation in the tree and all of the annotations that contain that annotation.

A subject heading is a library term chosen from a controlled list of annotations. A controlled list is a set of annotations that have been predetermined ahead of time. The annotator may not make up new subject headings. An object may have as many subject headings as desired by the annotator.

Works are annotated with two classifications, the Library of Congress Classification (LCC) and the Dewey Decimal Classification. (Tables 2 and 3 show the top LCC and DDC groups in LibraryThing.) A work has a position in both classifications. LCC and DDC encode their hierarchy information in a short string annotating a work, for example, GV735 or 811 respectively. The number 811 encodes that the book is about “Language and Literature” because it is in the 800s, “American and Canadian Literature” because it is in the 810s, and “Poetry” most specifically, because it is in the 811s. Likewise, “GV735” is about “Recreation and Leisure” because it is in GV, and “Umpires and Sports officiating” because it is in GV735. One needs a mapping table to decode the string into its constituent hierarchy information.

We flatten LCC and DDC for this paper, for example, 811 is treated as three groups {800, 810, 811}. This is, in some sense, not fair to LCC and DDC because the hierarchy provides additional information. However, we also ignore significant strengths of tagging in this work, for example, its ability to have thousands of annotations for a single work,
or its ability to show gradation of meaning (e.g., a work 500 people tag “fantasy” may be more classically “fantasy” than a work that only 10 people have tagged). In any case, the reader should recognize this difference between hierarchies and other library terms, and should note that our group model does not fully model hierarchical information.

Works are also annotated with one or more Library of Congress Subject Headings (LCSH). Each LCSH annotation is structured as one “LCSH main topic” and zero or more “LCSH subtopics” each selected from a vocabulary of phrases. For example, a book about the philosophy of religion might have the heading “Religion” (Main Topic) and “Philosophy” (Subtopic). We construct three sets of groups from the LCSH main and subtopics as follows:

**Set 1** We take all of the LCSH main topics and subtopics together to be one single group.

**Set 2** We take each (main topic, subtopic) pair together to be a group.

**Set 3** We take all main topics and subtopics to be separate groups.

We do not use LCSH’s hierarchical structure.\(^1\)

In practice, books rarely have more than three to six LCSH headings due to originally being designed for card catalogs where space was at a premium. It is also common for only the most specific LCSH headings to be annotated to a book, even if more general headings apply. Table 4 shows top LCSH headings in LibraryThing.

A **MARC record** is a standard library record that contains library terms for a particular book. In addition to classifications and subject headings, it includes a short fixed length string which has some useful data, which we call MARC 008. This string includes whether the book is a biography, and if so what type, whether the book is fiction, and some other data. Table 5 shows the top MARC 008 terms in LibraryThing. We define LCC, DDC, LCSH, LMARC008 to be the library terms of LCC, DDC, LCSH, and MARC008, respectively.

### 3. RESEARCH QUESTION

In this paper, we test whether social cataloging systems have seven features that we believe should be emulated from library systems:

1. **Objective, content-based annotations.** We believe that works should be organized objectively based on their content. For example, it makes sense to browse a system which is made up of groups of works with names like “History” and “Biography,” but it would be a nightmare to browse a system made up of groups of works with names like “sucks” and “my stuff” (Section 5.1).

2. **Appropriate group size frequencies.** A system made up of groups of works where each group contains two works would be difficult to browse, as would a system where all groups are made up of a million works. A system should have the right distribution of these group sizes in order to be usable (Section 5.2).

3. **Good coverage of the same groups as the library terms.** We believe that after decades of consensus, libraries have roughly the right groups of works. A system which attempts to organize works should end up with similar groups to a library (Section 5.3).

4. **Good recall.** A system should not only have the right groups of works, but it should have enough works annotated in order to be useful. For example, a system with exactly the same groups as libraries, but with only one work per group (rather than, say, thousands) would not be very useful (Section 5.4).

5. **Little synonymy in annotations.** There should not be multiple places to look for a particular object. This means that we would prefer tags not to have synonyms. When a tag does have synonyms, we would prefer one of the tags to have many more objects annotated with it than the others (Section 5.5).

6. **Consistent cross-system annotation use.** Across the same type of system, in this case, across tagging systems, we would like to see the systems use the same vocabulary of tags because they are annotating the same type of objects—works (Section 5.6).

7. **Consistent cross-system object annotation.** We would like the same work in two different tagging systems to be annotated with the same, or a similar distribution, of tags (Section 5.7).

These features each help to answer our research question: **how well do tags organize data?**

### 4. DATASETS

We use a dump of Library of Congress MARC records from the Internet Archive as the source of our library terms. We chose to use only those 2,218,687 records which had DDC and LCC library terms as well as an ISBN (a unique identifier for a book). We also use a list of approximately 6,000 groups in LCC from the Internet Archive, and a list of

<table>
<thead>
<tr>
<th>LCC</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>29211</td>
</tr>
<tr>
<td>History</td>
<td>22977</td>
</tr>
<tr>
<td>Biography (Subtopic)</td>
<td>14673</td>
</tr>
<tr>
<td>20th Century</td>
<td>11721</td>
</tr>
<tr>
<td>Fiction (Form of Work)</td>
<td>11214</td>
</tr>
<tr>
<td>Fiction (Subtopic)</td>
<td>10651</td>
</tr>
<tr>
<td>Juvenile Literature</td>
<td>7635</td>
</tr>
<tr>
<td>History and Criticism</td>
<td>7235</td>
</tr>
<tr>
<td>Biography (Form of Work)</td>
<td>6638</td>
</tr>
<tr>
<td>Great Britain</td>
<td>6312</td>
</tr>
</tbody>
</table>

**Table 4: Top 10 LCSH Terms.**

<table>
<thead>
<tr>
<th>MARC008</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Fiction</td>
<td>253446</td>
</tr>
<tr>
<td>Has Illustrations</td>
<td>178249</td>
</tr>
<tr>
<td>Has Bibliography</td>
<td>151391</td>
</tr>
<tr>
<td>Fiction</td>
<td>55812</td>
</tr>
<tr>
<td>Has Maps</td>
<td>25558</td>
</tr>
<tr>
<td>Juvenile Audience</td>
<td>24787</td>
</tr>
<tr>
<td>Has Plate Illustrations</td>
<td>16216</td>
</tr>
<tr>
<td>Individual Biography</td>
<td>12338</td>
</tr>
<tr>
<td>State Government Publication</td>
<td>9084</td>
</tr>
<tr>
<td>Autobiography</td>
<td>8423</td>
</tr>
</tbody>
</table>

**Table 5: Top 10 MARC 008 Terms.**

---

\(^1\)LCSH has some hierarchical structure (i.e., LCSH main topics can be broader than or narrower than other main topics). However, we do not use this structure because it is not always true that if a main topic \(mt_1\) is broader than another main topic \(mt_2\) that \(mt_1\) is broader than another main topic \(mt_3\) that \(mt_1\) is broader than \(mt_3\).
approximately 2,000 groups in DDC from a library school board in Canada as mapping tables for LCC and DDC.

We started crawling LibraryThing in early April 2008, and began crawling Goodreads in mid June 2008. In both cases, our dataset ends in mid-October 2008. We crawled a sample of works from each site based on a random selection of ISBNs from our Library of Congress dataset. LibraryThing focuses on cataloging books (and has attracted a number of librarians in addition to regular users), whereas Goodreads focuses on social networking (which means it has sparser tagging data). We gathered synonym sets (see Section 5.5) from LibraryThing on October 19th and 20th.

We use two versions of the LibraryThing dataset, one with all of the works which were found from our crawl, and one with only those works with at least 100 unique tags. The former dataset, which we call the “full” dataset, has 309,071 works. The latter dataset, which we call the “min100” dataset, has 23,396 works. We use only one version of our Goodreads dataset, a version where every work must have at least 25 tags.

5. EXPERIMENTS

In these experiments, we address the features described in Section 3. Each experiment begins with a short high level summary of the results, possibly has a brief preliminaries section, and then goes into detail about background, methodology, and the outcome.

5.1 Objective, Content-based Groups

Summary

Result 1: Most tags in both of our social cataloging sites were objective and content-based. Not only are most very popular tags (oc(t) > 300) objective and content-based, but so are less popular and rare tags.

Conclusion: Most tags, rather than merely tags that become very popular, are objective and content-based, even if they are only used a few times by one user.

Preliminaries: Tag Types

We divide tags into six types:

Objective and Content-based Objective means not depending on a particular annotator for reference. For example, “bad books” is not an objective tag (because one needs to know who thought it was bad), whereas “world war II books” is an objective tag. Content-based means having to do with the contents of the book (e.g., the story, facts, theme, genre). For example, “books at my house” is not a content-based tag, whereas “bears” is.

Opinion The tag implies a personal opinion. For example, “sucks” or “excellent.”

Personal The tag relates to personal or community activity or use. For example, “my book”, “wishlist”, “mike’s reading list”, or “class reading list”.

Physical The tag describes the book physically. For example, “in bedroom” or “paperback”.

Acronym The tag is an acronym that might mean multiple things. For example, “sf” or “the”.

Junk The tag is meaningless or indecipherable. For example, “b” or “jiowefijowef”.

<table>
<thead>
<tr>
<th>Tag Type</th>
<th>LibraryThing</th>
<th>Goodreads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective, Content of Book</td>
<td>60.55</td>
<td>57.10</td>
</tr>
<tr>
<td>Personal or Related to Owner</td>
<td>6.15</td>
<td>22.30</td>
</tr>
<tr>
<td>Acronym</td>
<td>3.75</td>
<td>1.80</td>
</tr>
<tr>
<td>Unintelligible or Junk</td>
<td>3.65</td>
<td>1.00</td>
</tr>
<tr>
<td>Physical (e.g., “Hardcover”)</td>
<td>3.55</td>
<td>1.00</td>
</tr>
<tr>
<td>Opinion (e.g., “Excellent”)</td>
<td>1.80</td>
<td>2.30</td>
</tr>
<tr>
<td>None of the Above</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>No Agreement by Annotators</td>
<td>20.35</td>
<td>14.30</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 6: Tag Types for Top 2000 LibraryThing and Top 1000 GoodReads Tags as Percentages.

Figure 1: Conditional Density Plot Showing Probability of a Tag Being Objective, Content-based, Being a Different Type, or not Having a Majority of Annotators Agreeing.

Preliminaries: Mechanical Turk

Amazon’s Mechanical Turk is a marketplace made up of requesters and workers. The requesters provide a task and set a price. The workers accept or decline the task. A task is a unit of work, like determining the type of a tag. We say a worker provides a determination of the answer to a task (for example, the tag “favorite” is an opinion). All of our tasks on Mechanical Turk were completed by at least five workers. We say that there was no majority if there was no answer to a task that was agreed upon by the majority of workers. We say the inter-annotator agreement is the pair-wise fraction of times two workers provide the same answer.

Details

If a tagging system is primarily made up of objective, content-based tags, then it is easier for users to find objects. In a library system, all annotations are objective and content-based in that they do not depend on reference to the annotator, and they refer to the contents of the book.

To produce an unbiased view of the types of tags in our sites, we used Mechanical Turk. We submitted the top 2,000 LibraryThing tags and top 1,000 Goodreads tags by annotation count to be evaluated. We also sampled 1,140 LibraryThing tags, 20 per rounded value of log(oc(t)), from 2.1 to 7.7. Overall, 126 workers examined 4,140 tags, five times per tag, leading to a total of 20,700 determinations. The inter-annotator agreement rate was about 65 percent.

Table 6 shows the top tag type results for LibraryThing and Goodreads. The results show that regardless of the
site, a majority of tags tend to be objective, content-based tags. In both sites, about 60 percent of the tags examined were objective and content-based. Interestingly, Goodreads has a substantially higher number of “personal” tags than LibraryThing. We suspect that this is because Goodreads calls tags “bookshelves” in their system.

Even if we look at tags ranging from $oc(t_i) = 8$ to $oc(t_i) = 2208$, as shown in Figure 1, the proportion of objective, content-based tags remains very high. That figure shows the probability that a tag will be objective and content-based conditioned on knowing its object count. For example, a tag annotating 55 objects has about a 50 percent chance of being objective and content-based. This suggests that (a) users are choosing objective, content-based tags on their own, and (b) tags which many users use are slightly more likely to be objective, content-based because users are less likely to overlap on non-objective or more personal tags.

5.2 Appropriate Distribution of Group Sizes

Summary

Result 2: Tags define a distribution of group sizes similar in shape to library systems.

Conclusion: Tags do a good job of creating appropriately sized groups for a collection.

Details

In this experiment, we look at the frequencies\(^2\) of groups in four different distributions:

\[
\begin{align*}
\{oc(l,i)|i \in L_{\text{LCC}}\} & \text{ LCC group sizes.} \\
\{oc(l,i)|i \in DDC\} & \text{ DDC group sizes.} \\
\{oc(l,i)|i \in L_{\text{LCSH}}\} & \text{ LCSH group sizes.} \\
\{oc(l,i)|i \in L\} & \text{ Combined library term group sizes.}
\end{align*}
\]

We compare these group frequencies to the frequencies of \(\{oc(t_i)|t_i \in T\}\), the tag group sizes.

We also consider “normalizing” the tagging distribution using sampling. The idea behind this normalization is that Section 5.1 showed that \(\approx 60\) percent of tags are objective and content-based (Table 6) and that this proportion is fairly constant across the size of groups (Figure 1). When we normalize, we sample 60 percent of the tags in \(T\) (the set of all tags) as \(T_{\text{norm}}\), and then use the distribution \(\{oc(t_i)|t_i \in T_{\text{norm}}\}\). (We cannot evaluate the hundreds of thousands of tags in our dataset, so this normalization approximates by randomly assigning 60 percent of tags to be objective and content-based.) This gives us an idea of what the counts of the objective, content-based tags might look like as opposed to all tags.

We use the “full” and “min100” LibraryThing datasets for this analysis. For each dataset, and for each (library term, tag) pair, we plot two sets of points on the same plot. A point at \((x,y)\) means that there are \(y\) groups which contain exactly \(x\) objects in the dataset.\(^3\) In each plot, one set of points are the group sizes and frequencies for a distribution of library terms, and the other set of points are the group sizes and frequencies for (normalized or unnormalized) tags.

Figures 2, 3 and 4 show tags versus library terms in five cases. In Figure 2(a), we can see that tags are similarly distributed to LCC, though the actual group sizes are somewhat different. In Figure 2(b), we see that tags seem even more closely distributed to LCSH than to LCC, though the group sizes are still different. In Figure 3(a), the tag distribution is compared to the combined distribution of all library terms. Here, the distributions are much closer than they were in either the LCSH or LCC cases. In Figure 3(b), we see that the normalized tag distribution and the combined distribution are almost exactly the same, both in distribution shape and group size frequency. The only real difference is in the highest object count terms (i.e., > 100,000). However, this extreme similarity disappears if we look at the “min100” dataset in Figure 4. In “min100,” each object tends to have many more tags on average than in the “full” dataset. This results in the group size frequencies being quite different, even though the distribution shape remains similar.

5.3 Coverage

Summary

Result 3: Many top tags have equivalent library terms. Tags contain more than half of the tens level DDC headings. There is a corresponding LCSH heading for more than 65 percent of top objective, content-based tags.

Conclusion: Top tags often correspond to library terms.
One of our goals is to look for pairs \((l_i, t_j)\) such that \(t_j \supseteq l_i\) or \(t_j \supseteq l_i\). However, evaluating whether \(t_j = l_i\) or \(t_j \supseteq l_i\) is expensive because it involves humans. Thus, we use data mining (specifically, association rules) to identify \((l_i, t_j)\) pairs where it is likely that \(t_j = l_i\) or \(t_j \supseteq l_i\).

An association rule is a relationship between two groups \(g_1\) and \(g_2\) of the form \(g_1 \rightarrow g_2\) (written \(g_1 \rightarrow g_2\)). In this work, we look only at association rules of the form \(l_i \rightarrow t_j\) (because we want to know when tags contain library terms, and because we want to experiment with recall, below). In this context, an association rule has a confidence which is equal to

\[
\text{conf}(l_i, t_j) = \frac{|O(l_i) \cap O(t_j)|}{|O(l_i)|}
\]

and a support equal to

\[
\text{support}(l_i, t_j) = |O(l_i) \cap O(t_j)|
\]

and an interest equal to

\[
\text{interest}(l_i, t_j) = \frac{\text{conf}(l_i, t_j) \cdot |O(t_j)|}{|O(l_i)|}
\]

A high confidence association rule \(l_i \rightarrow t_j\) usually means that either (a) \(l_i = t_j\), (b) \(t_j \supseteq l_i\) (most likely), (c) \(l_i \supseteq t_j\), or (d) \(l_i\) and \(t_j\) are related, but not by equivalence or containment. For example, the LCSH heading “Detective and mystery stories” is contained by the tag “mystery/detective” and the corresponding rule has high confidence. However, our highest confidence association rule involving “mystery/detective” has the LCSH heading “Horse racing” more specifically “fiction” because there are many horse racing mysteries. Association rules help us discover containment and equivalence between tags and library terms, but they are not accurate enough to be our sole tool. In our experiments, we use Mechanical Turk and manual evaluation as additional tools for evaluating association rules. We call the process of determining the type of relationship a rule represents evaluation. Evaluating high confidence association rules, rather than arbitrary pairs, helps us avoid misunderstanding the nature of how a tag is used. In a vacuum, a tag could have many meanings, but we at least guarantee through association rules that when making an evaluation, we do so based on both the semantics of the tag and the objects contained (rather than just the former).

**Details**

In this experiment, we ask whether tags provided by users give good coverage of library terms. In other words, do tags correspond to many of the annotations which have been determined to be important by expert consensus? Looking back at Tables 1, 2, 3, 4 and 5, we can see that the top tags all have library term equivalents. But what about tags that are not in the top 15?

We focus on 738 tags from the top 2000 LibraryThing tags which were unanimously considered objective and content-based in Section 5.1. These 738 tags are present in about 35% of the total tag annotations in our dataset. We gave the workers the top ten association rules \(l_i \rightarrow t_j\) for each of the top 738 tags \(t_j\), ranked by:

\[
\text{interest}(l_i, t_j) \times |O(l_i)|
\]

In practice, many tags had less than 10 rules, so the total number of rules given to workers was a little more than 5000.

Due to the difficulty of this task, agreement rate ranged between 34 and 68 percent depending on assumptions about agreement. We eliminated several bad users, and then split the rules into two sets: \(\text{set}_{201}\) where the workers evaluated that the tag contained the library term and \(\text{set}_{201}\) where the workers did not. \(\text{set}_{201}\) was constructed by taking all cases where a majority of the annotators said that the relationship was either \(l_i = t_j\) or \(t_j \supseteq l_i\). Finally, we manually examined all \(\approx 5,000\) rules. During this manual examination, we switched \(\frac{627}{1001}\) rules from \(\text{set}_{201}\) to \(\text{set}_{201}\), we switched \(\frac{449}{1989}\) rules from \(\text{set}_{201}\) to \(\text{set}_{201}\), and we marked equivalent rules.

This is naturally an imperfect process—what annotation contains another is a matter of both opinion and perspective. Nonetheless, we feel that the question of how tags relate to traditional library tools is important and that we should try our best to resolve it.

In terms of coverage, our methodology above produced 2924 rules where the tag contains the library term, and 524...
rules where the tag is equivalent to the library term. There are 450 unique tags in the “containment” rules, and 373 unique tags in the “equivalence” rules. More than half of the tags have directly equivalent library terms. This is especially surprising given that (a) we did not check all association rules, only the top 10 for each tag, (b) we know of at least some equivalencies that were not found due to our methodology, and (c) many of the tags are things you might not expect to find as library terms, like “suffering.”

Table 7 shows coverage of DDC by tags, either through equivalence or containment. (We use DDC for ease of exposition, but coverage of LCC is equally good if not better.) We give three measures of how well the hundred, ten, and one levels are covered by tags.

**Containment (Contained)** What proportion of the DDC groups are contained by a tag?

**Containment (Exact)** What proportion of the DDC groups have a tag which exactly contains the group and not a broader group?

**Equivalence** What proportion of the DDC groups have a tag which is equivalent?

For example, the first row, second column says $\frac{65}{100}$ ten level groups are contained by a tag which contains either a hundred level group or a ten level group. The second row, third column says that $\frac{178}{450}$ one level groups have a tag which contains exactly that group and not a larger group. The third row, first column says that $\frac{1}{16}$ DDC hundred level groups has an equivalent tag.

Almost half of the tags have an equivalent LCSH heading. We further analyzed LCSH, by looking for headings which exactly matched remaining tags without equivalencies, but were not discovered or evaluated above. We found a further 130 of the tags had LCSH headings, but we had not discovered them through our methodology. Overall, 503 of 738 tags had equivalents, or about 68 percent.

### 5.4 Recall

**Summary**

**Result 4:** Recall is low (10 to 40 percent) on the full dataset. Recall is high (60 to 100 percent) when we focus on popular objects (min100).

**Conclusion:** Tagging systems provide good recall for objects that are popular, but not necessarily for unpopular objects.

**Preliminaries: Recall**

After association rules have been evaluated as in Section 5.3, we know some groups that contain or are equivalent to other groups. If a rule $l_i \rightarrow t_j$ means $l_i = t_j$, then we say $l_i \in E(t_j)$. If a rule $l_i \rightarrow t_j$ means $t_j \supseteq l_i$, then we say $l_i \in C(t_j)$.

These metrics give us a way to determine how well a tag $t_i$ annotates the objects it is supposed to according to our library terms. We say the potential object set for a tag based on its contained or equivalent library terms is:

$$P_{t_i} = \bigcup_{t_i \in E(t_j) \cap C(t_j)} O(t_j)$$

We say recall is equal to:

$$\text{recall}(t_i) = \frac{|O(t_i) \cap P_{t_i}|}{|P_{t_i}|}$$

and that the Jaccard similarity between the potential object set and the objects contained by a tag is:

$$J(O(t_i), P_{t_i}) = \frac{|O(t_i) \cap P_{t_i}|}{|O(t_i) \cup P_{t_i}|}$$

**Details**

In this experiment, we use the association rules and evaluations from Section 5.3 to ask whether the tags provided by users have good recall of their contained library terms. An ideal system should have both good coverage (see Section 5.3) and high recall of library terms.

Figure 5 shows the recall for tags with at least one contained term on the full and min100 datasets. Figure 5(a) shows that on the full dataset, most tags have 10 to 40 percent recall. Figure 5(b) shows recall on the “min100”
dataset. We can see that when we have sufficient interest in an object (i.e., many tags), we are very likely to have the appropriate tags annotated. Recall is often 80 percent and up. Lastly, Figure 6 shows the distribution of Jaccard similarity between $O(t_i)$ and $P_t$. For most tags, the set of tag annotated works is actually quite different from the set of library term annotated works, with the overlap often being 20 percent of the total works in the union or less.

5.5 Synonymy

Summary

Result 5: Most tags have few or no synonyms. In a given set of synonyms, one tag is usually much more used than its alternatives.

Conclusion: Synonymy is not a major problem for tags.

Preliminaries: Synonymy

A group of users named combiners are users who mark tags as equivalent. (These are regular users of LibraryThing, not people working for us.) We call two tags that are equivalent according to a combiner synonyms or synonymous. A set of synonymous tags is called a synonym set. We write the synonym set of $t_i$, including itself, as $S(t_i)$. We calculate the entropy $H(t_i)$ (based on the probability $p(t_i)$ of each tag) of a synonym set $S(t_i)$ as:

$$p(t_j) = \frac{oc(t_j)}{\sum_{t_k \in S(t_i)} oc(t_k)} \quad H(t_i) = -\sum_{t_k \in S(t_i)} p(t_k) \log_2 p(t_k)$$

$H(t_i)$ measures the entropy of the probability distribution that we get when we assume that an annotator will choose a tag at random from a synonym set with probability in proportion to its object count. For example, if there are two equally likely tags in a synonym set, $H(t_i) = 1$. If there are four equally likely tags, $H(t_i) = 2$. The higher the entropy, the more uncertainty that an annotator will have in choosing which tag to annotate from a synonym set, and the more uncertainty a user will have in determining which tag to use to find the right objects.

Details

Due to the lack of a controlled vocabulary, tags will inevitably have synonymous forms. The best we can hope for is that users ultimately “agree” on a single form, by choosing one form over the others much more often. For example, we hope that if the tag “fiction” annotates 500 works about fiction, that perhaps 1 or 2 books might be tagged “fiction-book” or another uncommon synonym. For this experiment, we use the top 2000 LibraryThing tags and their synonyms.

Most tags have few, or no synonyms. However, there is a power law in the size of synonym sets (shown in Figure 7), and some tags have many synonyms. The largest synonym set is 70 tags (synonyms of “19th century”). Unlike one might expect, $|S(t_i)|$ is not strongly correlated with $oc(t_i)$ as shown in Figure 8. (Kendall’s $\tau \approx 0.208$.)

Figure 9 is a histogram of the entropies of the top 2000 tags, minus those synonym sets with an entropy of zero. In 85 percent of cases, $H(t_i) = 0$. The highest entropy synonym set, at $H(t_i) = 1.56$ is the synonym set for the tag “1001bymrbfd,” or “1001 books you must read before you die.” Less than fifteen tags (out of 2000) have an entropy above 0.5.

The extremely low entropies of most synonym sets suggest that most tags have a relatively definitive form. However, it should be noted that these synonym sets are somewhat strict. For example, “humor” and “humour” are left separate, because the former implies American wit and humor, while the latter implies British humor.

5.6 Consistent Cross-System Annotation Use

Summary

Result 6: The top 500 tags of LibraryThing and Goodreads have an intersection of almost 50 percent.

Conclusion: Similar systems have similar tags, though tagging system owners should encourage short tags.

Preliminaries: Information Integration

Federation is when multiple systems share data in a distributed fashion. Federation allows multiple sites to combine their collections, ultimately making the system more scalable. Information integration is the process of combining, de-duplicating, and resolving inconsistencies in the shared data. Two useful features for information integration are consistent cross-system annotation use and consistent cross-system object annotation. We say two systems have consistent annotation use if the same annotations are used overall in both systems. We say two systems have consistent object annotation if the same object in both systems is annotated similarly. Libraries achieve these two features through “authority control” (the process of creating controlled lists of headings) and professional catalogers, but we wonder to what extent taggers will have consistent annotation use and object annotation across systems.
For both LibraryThing and Goodreads, we look at the top 500 tags by object count. Ideally, a substantial portion of these tags would be the same, suggesting similar tagging practices, and similar views on how to tag books. Obviously, differences in the works tagged in the two systems will lead to some differences in tag distribution. However, both are predominantly made up of popular books, large amounts of fiction, and some other general interest books. Thus, ideally, we would expect both systems to have a very similar makeup of the top 500 tags.

The top 500 tags in both LibraryThing and Goodreads are quite similar. The overlap between the two sets is 189 tags, or about 38 percent of each top 500 list. We can also match by determining if a tag in one list is in the synonym set of a tag in the other list. If we do this, the overlap is even higher—231 tags, or about 46 percent of each list. This jump in overlap given synonyms suggests that while the “combiners” probably are not increasing the usefulness of their own system very much, they are probably making it much easier to integrate their system with other systems.

Much of the failed overlap can be accounted for by noting that Goodreads tends to have more multi-word tags. Multi-word tags lead to less overlap with other users, and less overlap across systems. We compute the number of words in a tag by splitting on spaces, underscores, and hyphens. On average, tags in the intersection of the two systems have about 1.4 words. However, tags not in the intersection have an average of 1.6 words in LibraryThing, and 2.3 words in Goodreads. This implies that in order for tagging to be federated across systems users need to be encouraged to use as few words as possible.

While there are 231 tags in the overlap between the systems (with synonyms), it is also important to know if these tags are in approximately the same ranking. Is “fantasy” used substantially more than “humor” in one system? We computed Kendall’s $\tau$ rank correlation between the two rankings from LibraryThing and Goodreads of the 231 tags in the overlap. Overall, Kendall’s $\tau$ is about 0.44, which is quite high. This means that if we choose any random pair of tags in both rankings, it is a little over twice as likely that the pair of tags is in the same order in both rankings as it is that the pair will be in a different order.

### 5.7 Consistent Cross-System Object Annotation

#### Summary

**Result 7:** Duplicate objects across systems have low Jaccard similarity in annotated tags, but high cosine similarity.

**Conclusion:** Annotation practices are similar across systems for the most popular tags of an object, but often less so for less common tags for that object.

#### Details

Do works which are shared across systems end up with the same set of tags? In other words, does “Winnie-the-Pooh” have the same set of tags in LibraryThing and Goodreads? We limited our analysis to works in both LibraryThing and Goodreads, where Goodreads has at least 25 tags for each book. This results in 787 works. Ideally, for each of these works, the tags would be substantially the same, implying that given the same source object, users of different systems will tag it in a similar way.

Figure 10 shows the distributions of similarities of tag annotations for the same works across the systems. We use Jaccard similarity for set similarity (i.e., with each annotation counting as zero or one), and cosine similarity for similarity with bags (i.e., with counts). Jaccard similarity is essentially a measure of how many annotations are shared. Cosine similarity with counts is counting how much the main annotations overlap with each other, because the distributions are peaked.

In Figure 10(a), we see that the Jaccard similarity of the
tag sets for a work in the two systems is actually quite low, perhaps one to six percent. One might expect that the issue is that LibraryThing has disproportionately many more tags than Goodreads, and these tags increase the size of the union substantially. To control for this, in Figure 10(b), we take the Jaccard similarity of the set of tags for a work in both systems, but we only take the top 20 tags for each work. Nonetheless, this does not hugely increase the Jaccard value in most cases. Figure 10(c) shows the distribution of cosine similarity values. (We treat tags as a bag of words and ignore three special system tags.) Strikingly, the cosine similarity for the same work is actually quite high. This suggests that for the same work, the most popular tags are likely to be quite popular in both systems, but that overall relatively few tags for a given work will overlap.

6. RELATED WORK

Similarly to our first experiment, work by Golder and Huberman [4] was the first to try to come up with a list of common tag types in a tagging system (specifically, delicious). However, we believe our work is possibly the first to analyze how tag types change over the long tail of tag usage (i.e., are less popular tags used different from more popular tags?). Chi et al. [1] looked at the entropy of tagging systems, in an effort to understand how tags grow, and how the groupings of tags change over time and affect browsing. This is similar to our second experiment, though with different tools, goals, and assumptions about what constitutes a good tag distribution. Specifically, we look for the tag distribution to be similar to library terms, whereas Chi et al. look for a low entropy tag distribution. Halpin et al.’s [5] work similarly looks at the nature of tag distributions with information theoretic tools.

Some work, for example, DeZelar-Tiedman [3] and Smith [9] looks at the relationship between tagging and traditional library metadata, similar to our third and fourth experiments on coverage and recall. However, these works tend to look at a few hundred books at most, and focus on whether tags can enhance libraries. Also related to these experiments, there has been some work on association rules in tagging systems, including work by Schmitz et al. [8] and some of our previous work [6]. However, that work focused on prediction of tags (or other tagging system quantities). We believe our work is the first to look at the relationship between tags and library terms using association rules.

Related to our fifth experiment on synonymy, Clements et al. [2] use LibraryThing synonym sets to try to predict synonyms. Similar to our sixth experiment on information integration, Oldenburg et al. [7] looked at how to integrate tags across tagging systems, though that work is fairly preliminary (and focused on the Jaccard measure). That work also focuses on different sorts of tagging systems, specifically, social bookmarking and research paper tagging systems, rather than social cataloging systems.

We are unaware of other work attempting to understand how tagging works as a data management or information organization tool in a quantitative way. We believe that we are the first to try to do this by contrasting tagging to established library systems.

7. CONCLUSION

Through the medium of social cataloging, we have shown that tags do a remarkably good job of organizing data. In a span of only a few years, LibraryThing has grown to tens of millions of books, and the groups developed by taggers are quite close to the groups developed by professional taxonomists. This is a testament both to the taxonomists, who did a remarkable job of choosing consensus controlled lists and classifications to describe books, and to tags which are unusually adaptable to different types of collections.

In our experiments, we found that the majority of tags are objective and content-based, though many other types of tags exist and are prevalent. We also found that tags have group size frequencies that are similar to those of library terms, implying that they allow similar quality of browsing. We found that tags have good coverage of many of the same groups as library terms, implying that the taggers found the right ways to divide up the books in the system. We found that tagging systems have acceptable recall, but that recall is much better when considered only over popular objects. We found that synonymy is not a huge problem, as most synonym sets are very low entropy and have one dominant tag. Lastly, we found that it may be possible to integrate tags across multiple sites, because many of the top tags for a given type of object (e.g., books) are similar, and because objects will tend to receive the same most popular tags (though, the raw sets covering works may be quite different across sites).

All of this brings us closer to a theory of tagging, and a theory of what organization of data best allows users to browse massive datasets. We have provided a look at how tags actually organize data in practice, by showing similarities in how both tags and library terms organize data for large sets of books. This is the first step in determining whether tags are a good organizational, data management, tool, rather than merely an interesting linguistic or social phenomenon to study.

8. REFERENCES


