

The Quest for Quality Tags

Shilad Sen, F. Maxwell Harper, Adam LaPitz, John Riedl
GroupLens Research
University of Minnesota
5-208 EE/CS Building, 200 Union Street SE
Minneapolis, MN 55455 USA
{ssen,harper,riedl}@cs.umn.edu, alapitz@neuralprophecy.com

ABSTRACT

Many online communities use tags – community selected words or phrases – to help people find what they desire. The quality of tags varies widely, from tags that capture a key dimension of an entity to those that are profane, useless, or unintelligible. Tagging systems must often select a subset of available tags to display to users due to limited screen space. Because users often spread tags they have seen, selecting good tags not only improves an individual’s view of tags, it also encourages them to create better tags in the future. We explore implicit (behavioral) and explicit (rating) mechanisms for determining tag quality. Based on 102,056 tag ratings and survey responses collected from 1,039 users over 100 days, we offer simple suggestions to designers of online communities to improve the quality of tags seen by their users.

Categories and Subject Descriptors

H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces—*Collaborative computing*; H.5.2 [Information Interfaces and Presentation]: User Interfaces

General Terms

Design, Experimentation, Human Factors

Keywords

tagging, moderation, user interfaces

1. INTRODUCTION

Member contributions power many online communities. Users upload images to flickr, bookmark pages on del.icio.us, and author encyclopedia entries at Wikipedia. These member-maintained communities harness their users’ effort to amass collections of millions of pictures, articles, and bookmarks. Navigating these large collections can be difficult. How should a user on flickr go about finding a freely available,

high quality image of a marine iguana among the 65 million uploaded photos? Computer vision algorithms cannot yet do a good job of selecting photos based on the wide variety of image features that are of interest to people [5].

Tags – words or phrases that describe items – have emerged as a flexible, rich means to navigate these corpuses. Tagging systems draw on contributions from ordinary users to out-scale expert maintained taxonomies. For example, in 200 years of existence the Library of Congress has applied their expert-maintained taxonomy to 20 million books¹. In contrast, in just four years, flickr’s users have applied their ad hoc tagging vocabulary to over 25 million photographs [16].

The resulting system is powerful. The search for “marine iguana” in the Creative Commons section of Flickr returns 19 photos – several strikingly good – free for use with attribution. The only returned photo not of an iguana shows the house of Senator John Warner of Virginia, who was once married to Elizabeth Taylor, who appeared in a 1964 movie called “On the Trail of the Iguana”. Every other photo is found in the search because of a tag added by a Flickr user.

Tagging systems scale well, but contributions from non-experts may reduce the quality of a system’s vocabulary of tags. For example, in the online community we study in this paper, users find that only 21% of the tags are worthy of general display. Low quality tags cluttering an interface may be useless or worse, they may be misleading, inappropriate, or offensive. Good tags, however, can make a system better by tying entities to one another to enhance browsing or search, or may serve as a source of descriptive information.

The lack of quality control on displayed tags is particularly dangerous given the self-reinforcing nature of tagging vocabularies. Conformity theory predicts that the tags that users see from other users will influence the tags that they in turn assign [2]. Conformity has been observed in practice. Golder and Huberman [12] and Cattuto [6] independently show that tagging vocabularies reach a stable equilibrium: once a tag becomes popular it remains popular. Sen et al. show that users tend to create tags resembling other tags they see in the community [17]. Systems that can select good tags not only improve the experience of the user who sees the tags, they also encourage those users to create good tags in return.

Selecting the right tags for display can be challenging for a number of reasons. Unlike data rich entities such as web pages and wikipedia articles, tags usually consist of a single unstructured word. As we mentioned earlier, tag quality can be quite poor. Moreover, tagging systems do not have much

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2007 ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

¹<http://www.loc.gov/about/reports>

room for error. Due to limited screen space, many systems can only display a few tags from among the many users have applied. With all these challenges, how should del.icio.us select the five tags it displays for the for the website digg from among the 10,688 unique tags users have applied to it?

Our goal in this research is to understand methods for selecting high quality tags for display while suppressing low quality tags. We explore several lightweight interfaces for collecting member feedback about tags, and examine which interfaces lead to the richest metadata for understanding the quality of individual tags. We then develop several approaches for predicting tag quality based on either implicit system usage data or on explicit member feedback.

We structure our paper around five research questions. Our first two research questions explore the effects of the rating interface on the tags displayed in a system. Rating interfaces that evaluate tag quality based on explicit ratings can only be effective for those tags that have been rated. Our first research question examines the relationship between rating interface and rating quantity:

RQ1: Which rating interfaces lead to the most ratings?

Increased rating quantity is only valuable to the extent that it improves the tags displayed by a tagging system. Our second research question examines this relationship directly.

RQ2: Which tag rating interfaces should designers implement to better select the tags they show to users?

Of course, tag ratings do not inherently determine which tags are displayed - a system must implement a *tag selection method* drawing on both tag ratings and non-ratings tag data. Our remaining research questions explore three fundamental signals a tag selection method may use:

RQ3: Can we determine the tags a user wants to see based on other users' behavior?

RQ4: Can we determine the tags a user wants to see based on a user's own ratings?

RQ5: Can we determine the tags a user wants to see based on other users' ratings?

The rest of the paper is organized as follows. In section 2 we summarize existing research related to tag selection methods. Section 3 presents our tag rating implementation and describes our experimental setup. Section 4 discusses RQ1, which relates the tag rating interface to rating quantity. Sections 5, 6, and 7 explore RQ3, RQ4, and RQ5, which investigate specific tag selection methods. We conclude in section 8 with a return to RQ2, design implications, discussion, and several ideas for future research.

2. RELATED WORK

Although public bookmarking systems such as Fab [3], Knowledge Pump [11], and Pharos [4] have been available since the 1990's, Millen et al. point to tagging as a key reason current social bookmarking systems have enjoyed greater success [15]. As tagging systems became more popular, Shirky [18] was among the first of many bloggers and technology critics who argued that traditional controlled ontologies improperly describe the way in which information is now organized. Our work furthers early studies of tagging communities by analyzing the quality of tags created in an

online community.

In early academic research on tagging communities, MacGregor and McCulloch [14] explore the relative merits of controlled versus evolved vocabularies, arguing that evolved ontologies engage users but lack the precision of their controlled counterparts. Golder and Huberman indicate that the proportions of tags applied to a given item in del.icio.us appear to stabilize over time, and suggest that community members may be influenced by what they see [12]. Cattuto furthers their work by presenting a generative model for users' tagging that predicts the rate at which both particular users and entire communities re-use tags [6]. In earlier work, we show that the tags a user sees influence the tags they create themselves [17]. We also classify tags as generally factual, subjective, or personal (intended for the tag creator themselves), and find that users generally prefer factual tags over subjective tags and strongly dislike personal tags. Our research extends earlier work describing how users choose tags to the novel problem of how systems might select tags to show a user from among a large collection of tags that have been applied by other users.

Several researchers have studied moderation in online communities. Cosley et al. find that "Wiki-like" systems that immediately display user contributions lead to more contribution than systems that require members to review contributions before they are displayed [7]. In other work, Cosley et al. show that intelligent task routing can be used to help users find tasks they might complete to improve the system [8]. Lampe and Resnick analyze the moderation system utilized on the online forum slashdot² [13]. They find that although the community perceives that forum moderations are generally fair, comments that are assigned low scores, or posted late in a conversation are often overlooked by moderators. Arnt and Zilberstein explore machine learning techniques for predicting moderation scores in online forums [1]. Our research differs from the general work on moderation of contributions in that we focus on a type of contribution (tags), and investigate ways in which user interfaces may improve moderation.

3. METHODS

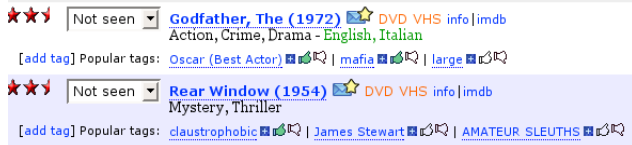
As a platform for our analyses, we used the MovieLens³ movie recommendation system. MovieLens members can tag movies, and use tags contributed by others in the community to find or evaluate movies. Since we introduced tagging features to MovieLens in January 2006, MovieLens users have created 52,814 tag applications resulting in 9,055 distinct tags. (A *tag* is a particular word or phrase used in a tagging system. A *tag application* is the result of associating a tag with a system entity.) 2,344 users have applied at least one tag (13.5% of active users). Further details of MovieLens and the MovieLens tagging system can be found in [17].

In order to study explicit tag feedback, we introduced tag ratings to the MovieLens community. Our design of a tag rating system was based on two guiding principles: users should be able to rate tags with a single click, and the ratings interface should require minimal screen space. Since a star-based rating system requires too much space, we selected a thumbs up / thumbs down rating system, similar to that

²<http://www.slashdot.org>

³name anonymized

Figure 1: Tags as they appear on the MovieLens search results screen, next to the experimental thumbs up and thumbs down ratings widgets.



used in many commercial applications such as Amazon⁴, TiVo⁵, and reddit⁶.

While many commercial applications incorporate both thumbs up and thumbs down ratings, several only employ one or the other. For example, BoardGameGeek originally employed thumbs up and down moderation, but shifted to only thumbs up moderation to “make it harder for people to “gang up”” and “reduce hurt feelings.”⁷ In sites such as YouTube, users provide positive feedback about items by marking items as “favorites.” Other sites allow solely negative feedback. Users of Google Video, for example, may mark tags as “spam” but have no means of providing positive feedback.

To investigate the utility of different rating interfaces, we randomly split users into four experimental groups representing possible combinations of positive (thumbs up) and negative (thumbs down) ratings widgets:

1. Control group **C** was not shown any tag rating widgets.
2. Group **U** was only shown the thumbs up tag rating widget.
3. Group **D** was only shown the thumbs down tag rating widget.
4. Group **UD** was shown both the thumbs up and thumbs down tag rating widgets.

The tag rating interface appeared alongside all tag applications appearing on the MovieLens search results page (Figure 1) and movie details page (Figure 2). Search results pages displayed up to three tags per movie, while the movie details page displayed up to twenty tags. MovieLens randomly selected and ordered tags for display from among the tags applied to a movie.

To help motivate users to provide tag ratings, we implemented simple user interface responses to rating actions. Tags shift to the front of a movie’s tag list in response to a positive rating, and tags move to the end of the movie details page list and are hidden from the search results page in response to a negative rating. We incorporated AJAX javascript controls to enable fast, lightweight rating interactions. We enabled the tag rating features on January 21, 2007 and collected data for one hundred days.

While the thumbs up and down ratings provided coarse data about tag quality, we also wanted a “gold standard” data set for evaluating our techniques for selecting tags to display. To this end, we emailed 2,531 active MovieLens

⁴ Amazon.com uses thumb ratings for meta-reviewing.

⁵ The TiVo digital video recorder collects user feedback through a thumb-based interface

⁶ The news aggregation service reddit.com allows users to click an up-arrow or a down-arrow for each article.

⁷ <http://www.boardgamegeek.com/thread/156510>

Figure 2: Tags and the experimental ratings widgets as they appear on the movie details screen.

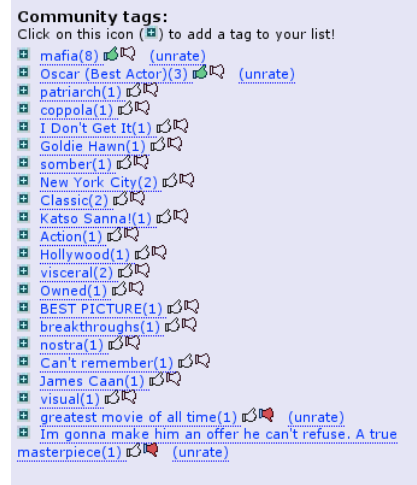


Figure 3: We asked users to rate tags for five movies on a one-to-five scale. We instructed them that MovieLens would only choose to show them tags rated 3, 4, or 5 stars.

tag	don't show tag		show tag		
	★	★★	★★★	★★★★	★★★★★
Oscar Winner	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
twist ending	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
seen at the cinema	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
kevin spacey is soze	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
mindfuck	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
imdb top 250	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

users and asked them to complete an online survey in which they provide feedback on tag quality. Users were asked to rate up to twenty tags applied to five movies on a five star scale. The five selected movies consisted of:

- Two movies that users most frequently rated and tagged (*The Usual Suspects* and *Star Wars Episode IV - A New Hope*).
- Two randomly-selected, frequently-tagged movies that the user had rated.
- One randomly-selected, frequently-tagged movie that the user had not rated.

Figure 3 shows an example screen from the survey. As a point of reference, users were instructed that MovieLens would only display tags rated 3,4, and 5 stars. 577 users responded to the survey (22.8% response rate) and rated at least one tag application. 546 users rated tags for all five movies. We gave users the option of continuing to rate tags after they completed rating tags for their first five movies.

Users provided 74,987 one-to-five ratings. Two users provided more than 1,000 ratings, while 253 users provided 100 or more ratings. The distribution of the one-to-five ratings is shown in Table 1. Users deemed 38% of rated tag applications worthy of display. The average tag rating was 2.17.

Other than the tag applications themselves, we base our analyses on three types of data. We study whether patterns in aggregate user behavior, such as searches for tags, indicate that tags should be displayed or hidden. We use thumbs up

Table 1: Distribution of one-to-five tag ratings by rating value. The average overall rating is 2.17

tag rating value	1	2	3	4	5
percentage of ratings	46%	14%	21%	10%	7%

and thumbs down tag ratings to evaluate the relative utility of different rating interfaces, and explore their predictive power for tag selection methods. The one to five star survey ratings serves as a “gold standard” for evaluating selection methods and to better understand users liking for tags.

4. RATING INTERFACES

We have argued that up/down ratings interfaces are prevalent in modern systems, and we believe they are an appropriate light-weight interface for soliciting feedback about tags (and other ubiquitous system entities). One decision that a designer of such a ratings system faces is whether to include both a positive and a negative ratings widget, or if one or the other alone will provide sufficient data to build accurate models of quality. Designers of commercial systems are divided on this issue, even within the same domain: the social news site reddit⁸ has both up and down arrows, whereas the social news site digg⁹ has a “digg” button in a highly visible place on the interface, with a less visible “bury” button elsewhere. In this section, we investigate this decision by examining data collected in a field study of several variants of a tag ratings system in MovieLens.

Methods for selecting tags to display depend on data – either implicit data about user behavior, or data collected explicitly from users. Many explicit ratings-based systems find collecting sufficient data a challenge. Thus, a key question is which interfaces attract the most ratings. Our first research question is:

RQ1: Which rating interfaces lead to the most ratings?

Table 2 shows a summary of up/down ratings applied during the experimental period by users in the different groups. In total, 460 users (7.3% of active users during the time period) generated a total of 27,069 tag ratings. 72% of tag ratings occurred from the search results page, while 28% occurred on the movie details page. A small number of users supplied the majority of tag ratings. For example, the top rater provided 10.4% of all tag ratings (2,823), and the top 20% of raters provided 93.5% of all tag ratings (25,322)¹⁰ 51.5% of raters applied 3 or fewer ratings.

Our first finding is that the presence of different ratings interfaces leads to significant differences in ratings contributions. The descriptive statistics from Table 2 give an intuitive feel for the results. Users in Group UD rated more times (13,841) than users in Group D (11,903) or in Group U (1,325). Also, more users in Group UD rated one or more times (14.2%) as compared with users in Group D (9.7%) or Group U (5.1%).

These differences are statistically significant. Because the distribution of work per-user is strongly skewed, we must apply non-parametric statistical tests to determine differ-

⁸ www.reddit.com

⁹ www.digg.com

¹⁰This distribution is common in member-maintained communities. For instance, in Wikipedia the most prolific 10% of users generate 80% of all edits [19].

ences. To measure the differences in per-user ratings between groups, we examine the ratings of all users who log in to the system during the experimental period. We test for differences using a one-way Wilcoxon test, and report significance based on the p-value resulting from a Chi-Square approximation. Users in Group UD rated more than users in Group D ($n = 3181$, means 8.65 vs. 7.53, $ChiSquare = 14.64$, $DF = 1$, $p < 0.001$), and they also rated more than users in Group U ($n = 3176$, means 8.65 vs. 0.84, $ChiSquare = 75.85$, $DF = 1$, $p < 0.001$). Users in Group D rated more than users in Group U ($n = 3157$, means 7.53 vs. 0.84, $ChiSquare = 25.04$, $DF = 1$, $p < 0.001$).

We also find that more users from Group UD contributed one or more tag ratings than from either of the other experimental groups. To test for significance, we conduct a likelihood ratio Chi-Square test. We find that users in Group UD were more likely to rate one or more tags than users in Group D (14.19% vs. 9.68%, $ChiSquare = 15.47$, $p < 0.001$), and they were also more likely to rate than users in Group U (14.19% vs. 5.08%, $ChiSquare = 78.45$, $p < 0.001$). Users in Group D were more likely to rate one or more tags than users in Group U (9.68% vs. 5.08%, $ChiSquare = 24.84$, $p < 0.001$).

Although on average, users in group D generated more negative ratings per-user as compared with users in group UD (means 7.53 vs. 6.13), this difference is not statistically significant using a Wilcoxon test ($n = 3181$, $ChiSquare = 0.05$, $df = 1$, $p = 0.82$). The difference in the means might be attributed to the presence of the most prolific rater in Group D, who singlehandedly rated 2,823 times.

Interestingly, we do find that users are more likely to rate tags positively in the presence of a thumbs-down rating widget. This is demonstrated by the fact that users in Group UD gave a thumbs up to an average of 2.5 tag applications, while users in Group U gave a thumbs up to just 0.8 tag applications. This difference is statistically significant, using a Wilcoxon test ($n = 3176$, $ChiSquare = 36.24$, $df = 1$, $p < 0.001$).

We thought the additional up ratings in the UD group might be due to tag “churn” introduced by negative tag ratings (negatively rated tags disappear and the user is presented with additional tags to rate). To test this hypothesis, we measured the tag-specific probabilities that a displayed tag would be rated positively across both Groups U and UD. We then calculated each group’s expected number of up ratings based on their displayed tags. We find that the number of up ratings in Group UD is 1.5 times the expected number, while the U group is half the expected number. Therefore, we cannot attribute the extra positive ratings in the Group UD to tag churn. Apparently there is something about the presence of both ratings in the interface that leads to more up ratings.

Overall, we find that the interface containing both up and down ratings widgets led to the greatest levels of contributions. We later return to the impact of these contributions on tag selection methods. However, the general message is that greater contributions leads to greater coverage, and therefore more successful interfaces for displaying high quality tags.

Our second finding is that users contributed more negative ratings than positive ratings, especially among users who rated more than three tags. Across all three experimental groups, we collected more than four times as many negative

Table 2: Statistics for each experimental group. The group with up and down ratings generated more positive ratings than the group with only up ratings. The number of raters in each group also varied significantly.

Group	Num Users	Num Raters	Thumbs Up	Thumbs Down	Total Thumbs
control (C)	1494	0 (0.0%)	0	0	0
up only (U)	1576	80 (5.1%)	1325	0	1325
down only (D)	1581	153 (9.7%)	0	11903	11903
up and down (UD)	1600	227 (14.2%)	4027	9814	13841
total	6251	460 (7.3%)	5352	21717	27069

ratings as positive ratings (21,717 vs. 5,352).

The difference in the quantity of down ratings versus up ratings is awkward to statistically verify. As stated above, Group UD rated more positively than Group U. We might therefore speculate that this difference is due to some aspect of the up-only interface which makes it less attractive to provide ratings. Therefore, we cannot fairly factor Group U into the comparison between up and down ratings. We are left with a paired Wilcoxon test among users of Group UD.

When looking only at Group UD, we find that a majority (119 vs. 101) of users actually rated more positively than negatively. The remainder (7 users) rated equal numbers up and down. There is no statistical difference across these users in per-user up ratings vs. down ratings using a Wilcoxon test ($n = 1600$, $W = 12$, $p = 0.99$).

However, when we look only at the 108 users in Group UD who have rated more than three times, we find significance ($n = 108$, $W = 522$, $p = 0.045$), accounting for the large overall difference in Table 2 (9,814 down ratings vs. 4,027 up ratings). We might state that committed raters contribute more down ratings than up ratings. However, as we discuss in section 8, this result is likely the result of the overall quality of tags in the MovieLens system, rather than the result of an innate preference for rating things down.

While “normal” tag raters produce similar quantities of positive and negative ratings, “power” tag raters strongly favor negative ratings. Differences in the number of positive or negative ratings may impact the effectiveness of certain tag selection methods. For instance, if users rate more negatively than positively, systems might be able to identify bad tags more easily than good ones. RQ2 directly explores the relationship between tagging interface and selection quality in the context of specific tag selection methods. We now move on to explore selection methods, but will return to RQ2 in section 8.

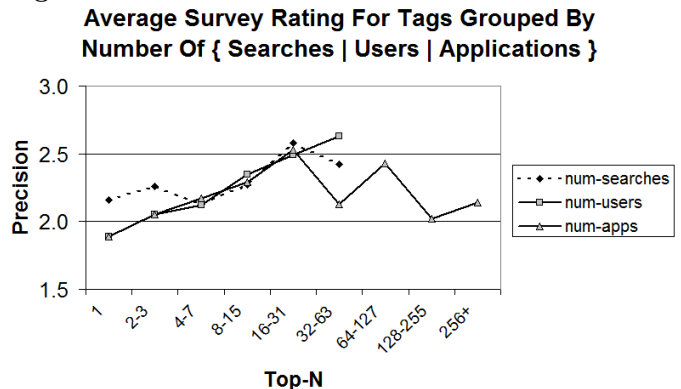
5. PREDICTING TAG QUALITY BASED ON AGGREGATE USER BEHAVIOR

Systems such as flickr and del.icio.us have attracted millions of users and generated vast amounts of behavioral data about the tags users created, searched for, and browsed. Ideally, designers of existing online tagging communities might estimate tag quality by analyzing existing behavioral data without having to collect explicit feedback about tags. In this section, we form predictions based on implicit measures of tag quality, such as the number of users who have applied a tag. We test those predictions against the gold standard of the user surveys in an effort to answer:

RQ3: Can we determine the tags a user wants to see based on other users’ behavior?

Perhaps users apply higher quality tags more often than

Figure 4: Average tag quality grouped by the number of tag applications, number of users who applied the tag, and number of users who searched for the tag.



low quality tags. If so, then the number of times a tag has been applied might be a reasonable proxy for its quality. A tagging system might wish to preferentially display tags applied many times, or hide tags that have been applied fewer than some minimum number of times. Motivated by this possibility, we examine the **num-apps** tag selection method which predicts tag quality based on the number of times a tag has been applied.

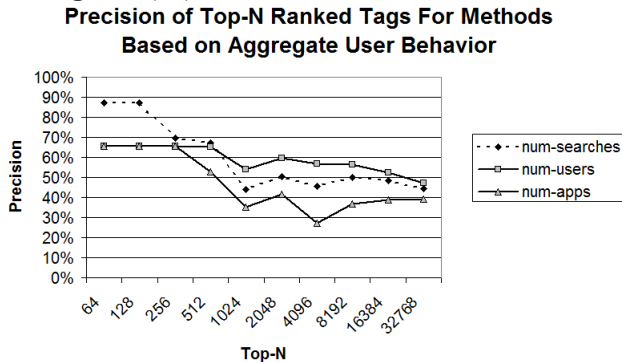
Figure 4 shows the average tag survey rating (a five star scale) grouped by the number of tag applications. Tags applied once have the lowest average rating (1.89), tags applied 16-31 times have the highest average rating (2.53), and the tags applied most often (256 or more times) have an average rating of 2.14.

We might expect the most often applied tags to be the highest rated, but this is not the case - users gave low average ratings for several of the most frequently applied tags. This may be attributable to an abundance of “personal” tags intended solely for their creator; four of the five most rated tags applied 256 or more times are personal: “dvd”, “own”, “seen at the cinema”, and “eric’s dvds.”

Personal tags appear to reduce the accuracy of the previous tag selection method. Sen et al. showed that personal tags are generally used frequently by only a few users [17]. Systems might show fewer personal tags by normalizing each user’s influence over the selection method. We now explore the **num-users** selection method, which predicts tag quality based on the number of *users* who have applied each tag.

Figure 4 shows the average tag rating grouped by the number of users who have applied the tag. The average rating of the tag used by the most users (32 or more) reaches

Figure 5: Precision of selection methods based on other users’ behavior top-n ranked survey ratings. Ratings of 3, 4, and 5 stars are viewed as desirable.



2.63 out of 5 stars. A clear upward trend is apparent: tags applied by more users are rated higher than tags applied by fewer users.

Perhaps users search for, and click on, good tags more frequently than bad tags. We analyzed 19,458 tag search and click events, and found that a few users who search for the same tag many times bias the number of searches per tag. For this reason, the **num-searches** selection method normalizes each user’s weight by focusing on the number of users who search for a tag.

Figure 4 shows average tag rating grouped by the number of users who clicked on each tag. As with the number of users who applied each tag, we see a gentle upward trend. Tags searched for by 16-31 users have an average rating of 2.58, while those searched for by 32 or more users have an average rating of 2.42. The small decline in average rating can be accounted for by two “personal” tags many users clicked on: “seen more than once”, and “erlend’s dvds.”

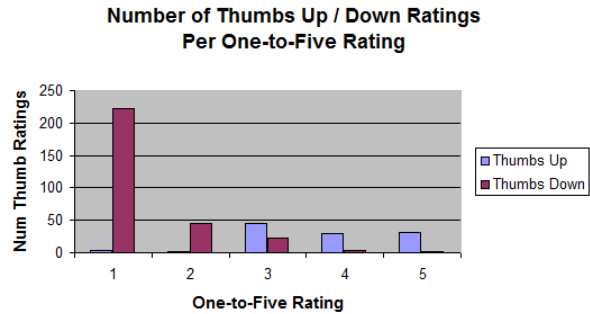
Unlike the first two selection methods, the search-based selection method can only generate predictions for tags users have searched for. The selection method achieves a prediction *coverage* of 70.2% of the tag survey ratings.

Average rating is just one possible measure of tag selection quality. Perhaps users wish to minimize the number of low-quality tags displayed. Inspired by this proposition, we ranked all survey ratings outputted by the previous three selection methods, and measured the precision (fraction of 3, 4, or 5 star ratings) at different thresholds. Figure 5 shows each selection method’s precision at different thresholds. We used a logarithmic scale on the x-axis to facilitate later comparisons with methods having lower coverage.

Num-searches performs well at the high end: 87% of the 128 tag survey ratings for the most searched-for tags were rated three, four or five stars. Num-users also performs consistently: more than half of the top ranked 16,384 survey ratings are rated three or higher. Num-apps, on the other hand, performs erratically.

The top-ranked precision numbers may seem higher than expected based on our earlier analysis using average survey ratings. For instance, while num-users achieves a precision of 56% for the top 8,192 survey ratings, the average of these ratings is only 2.63 (below the display threshold). This difference can be explained by the distribution of survey ratings: while ratings are divided relatively evenly among three, four, and five stars, there are far more one star ratings than

Figure 6: Mapping between thumbs up/down tag ratings and one-to-five survey ratings. Thumbs down ratings are mostly rated 1, while thumbs up ratings are evenly split between 3,4 and 5.



two star ratings. Low survey ratings affect survey averages disproportionately more than high survey ratings.

We find selection methods normalizing each user’s influence, such as num-users and num-search, to be more robust than methods which can be biased by a few power users (such as num-apps). Although the three selection methods appear to correlate with user liking for tags, none of them seem to be individually sufficient. Even in num-users, the overall top performer, users approve of barely half of the top ranked 22% of survey ratings (n=16384).

Although most real-world tagging systems do not have access to survey data as a gold standard, we believe that our use of them is justified. We hope that our conclusions will provide general insights into the way in which users evaluate tags. We also believe that many large tagging sites would eagerly conduct a small survey if it improved the quality of displayed tags.

6. PREDICTING TAG QUALITY BASED ON A USER’S OWN RATINGS

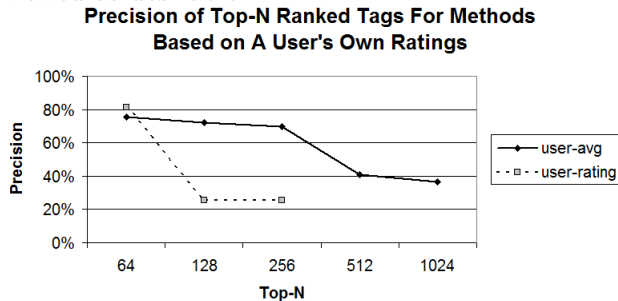
In the last section we showed that tag selection methods such as num-users and num-searches based on implicit behavior have some predictive power. As a more direct alternative, systems may use thumbs up or thumbs down feedback to select the tags a user wants to see. Research question 4 explores tag selection methods based solely on a user’s own ratings:

RQ4: Can we determine the tags a user wants to see based on a user’s own ratings?

In MovieLens, users rate specific applications of tags to movies. For instance, suppose Sally rates the tag “zombies” on the movie “28 Days Later” positively. Our first ratings-based tag selection method, **user-rating**, simply concludes that users like the tag *applications* they rate thumbs up and dislike the tag applications they rate thumbs down.

We begin by better understanding what Sally’s thumbs up rating for “zombies” on “28 Days Later” tells us about her survey rating (survey ratings use a higher-precision five star scale). Figure 6 shows the number of thumbs up and down ratings that were mapped to each survey rating. The meaning of thumbs up and down are clearly distinct. Down ratings map to a one or two star rating, while up ratings map to a rating of three stars or higher. While 80% of thumbs down ratings received a survey rating of 1 star, the thumbs up ratings were equally likely to be rated 3,4, or 5 stars.

Figure 7: Precision of selection method’s based on user’s own ratings. Ratings of 3, 4, and 5 stars are viewed as desirable.



The less extreme mapping for thumbs up ratings may be due to our annotation of the one-to-five survey scale: we told MovieLens users that 3, 4, and 5 star ratings would be displayed.

Figure 7 shows the precision of the application-based selection method. While the method achieves a precision of 81% for the top 64 survey ratings (those survey ratings also rated a thumbs up), it can only generate predictions for tag applications with an associated thumb rating. The tag selection method can only predict for one out of two hundred survey ratings, leading to a coverage of 0.5%.

Suppose that Sally rates the tag “zombies” on both “28 Days Later” and “Dawn of the Dead.” If Sally rates tags “zombies” consistently, her second rating of zombie may be wasting her valuable effort. To reduce Sally’s effort a system might assume that she will rate the tag “zombies” positively for all movies.

We evaluate this broader interpretation of tag ratings by measuring the effectiveness of a user’s average tag rating as a tag selection method (**user-avg**). We encoded thumbs up ratings as +1 and thumbs down ratings as -1 (we use this encoding throughout the rest of the paper). Tags with more than one user rating are weighted more heavily by adding one “neutrally” rated (0) tag to every average calculation (this is equivalent to using an uniform beta prior [10]). For instance, two positive and one negative rating will result in an adjusted rating of $\frac{0+1+1-1}{4} = 0.25$.

The average-based selection method achieves a precision of almost 70% for the top 256 survey ratings. The coverage improves by a factor of four to 1.9%, but still remains quite low.

In summary, a user’s tag ratings serve as strong predictors of their liking for particular tags. This precision comes at the expense of coverage - even if we extend user ratings of tag applications to apply to all occurrences of the tag we only cover 1.9% of survey ratings. Our results also indicate that systems may want users to rate tags instead of tag applications. The intra-rater reliability of the one to five star survey data also supports this conclusion; the average variance for a user’s rating of the same tag is only 0.175 on a five point scale.

7. PREDICTING TAG QUALITY BASED ON AGGREGATE USER’S RATINGS

Sally is not the only MovieLens user who likes the tag “zombies.” In fact, 81% of all thumb ratings for “zombies” in MovieLens are thumbs up ratings. If several raters agree

Table 3: Top 10 most controversial tags based on thumb ratings as measured by expected entropy (Appendix A).

tag	entropy	up	down
comedy	0.987	28	30
classic	0.986	25	24
stylized	0.983	20	21
nudity (full frontal)	0.980	18	20
romance	0.980	18	17
quirky	0.977	25	20
magic	0.974	18	15
animation	0.974	26	20
steven spielberg	0.973	12	12
sci-fi	0.972	14	17

on a tag’s quality, a system may be able to conclude that most users have similar opinions of the tag, increasing the tag selection method’s coverage to *all* users. In this section we explore how systems might select tags for display based on aggregate user thumb ratings.

RQ5: Can we determine the tags a user wants to see based on other users’ ratings?

To capture users’ aggregate tag opinions, we considered the **global-avg** selection method which ranks tags by their overall average rating across all users. As in the user-avg, we smoothed average ratings by adding a single neutral rating. Figure 9 shows that the precision for the highest ranked tags is slightly lower than selection methods based on a user’s own ratings. However, the decreased precision is offset by a 49x improvement in coverage to 93%.

Users obviously don’t agree on all tags. Table 3 lists the most controversial tags as measured by expected entropy (Appendix A). Controversial tags appeared to contain information that is already displayed in MovieLens (comedy, sci-fi, steven spielberg), subjective (classic, stylized, quirky), or about a controversial topic (nudity - full frontal).

We wondered whether certain types of tags lead to different levels of agreement across users. We discovered a difference in agreement for “good” and “bad” tags. We divided tags into those with one-to-five means above and below the 3 star display threshold, and measured the average variance across all users’ ratings for the tag. While the average variance for low-rated tags was 0.72, the average variance for highly-rated tags was 1.15. Users clearly agreed more about bad tags than good tags.

Perhaps Sally provided MovieLens’s fifth positive rating for “zombies” and no users had rated the tag negatively. This high level of initial agreement offers a promising signal for tag quality that can be easily implemented by system designers.

The previous four ratings for “zombie” may have all come from the same user (we know from section 6 that a user will generally rate a tag consistently). To be sure that the initial consecutive ratings are independent confirmation, a designer may want to require that they come from different users.

Based on these scenarios, we now examine the **consec-apps** selection method, which ranks tags based on the number of initial identical ratings, and the similar **consec-users**, which requires that the ratings come from different users.

We begin with an intuitive analysis of repeated ratings

Figure 8: Percent of remaining ratings that, after an initial number of identical ratings, remain positive or negative.

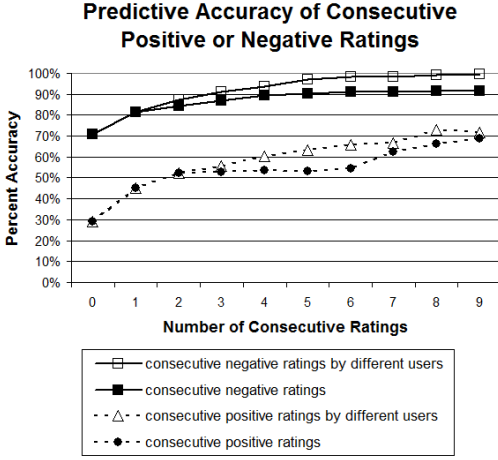
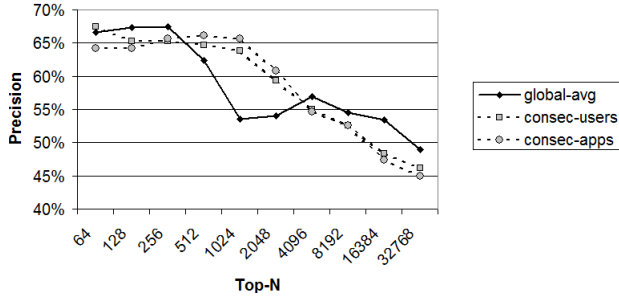


Figure 9: Precision of selection method’s based on other users’ ratings. Ratings of 3, 4, and 5 stars are viewed as desirable.

Precision of Top-N Ranked Tags For Methods Based On Aggregate User Ratings



that translates easily into system implementation. Figure 8 shows the percent of ratings that, after a certain number of initial consecutive positive or negative ratings, remain positive or negative. The graph presents both the count by-application and by-user metrics. Both metrics for negative consecutive ratings serve as accurate predictors. After a tag receives four consecutive thumbs down ratings (regardless of user), 90% of the remaining ratings will be thumbs down. On the other hand, even after a tag receives strictly thumbs up ratings by 9 different users, only 71% of the remaining tags are positive.

Both consec-apps and consec-users perform similarly on the rank / precision analysis that we used to evaluate prior selection methods. Consec-users yields a precision of 67% for the 64 top-ranked tags, compared to 64% for consec-apps. Both methods achieve over 50% accuracy for tags that were initially rated thumbs up.

In general, we find selection methods based on aggregate ratings to achieve slightly lower precision than methods based on a user’s own ratings, but with much higher coverage. The precision of the aggregate rating methods seems similar to the precision of methods based on aggregate implicit behavior. This does not mean that ratings-based methods do not provide additional benefit. If the im-

Table 4: Tag Selection Method’s Coverage of Tag Survey Ratings Per Experimental Group.

Method	C	U	D	UD
user-rating	0.0%	0.03%	0.7%	1.2%
user-avg	0.0%	0.4%	2.6%	4.6%
global-avg	0.0%	37.8%	77.2%	88.0%
consec-users	0.0%	37.8%	77.2%	88.0%

PLICIT and explicit selection methods excel at different types of tags, systems may draw on both methods to construct a more accurate hybrid selection method. We investigate one such method in the next section.

8. DISCUSSION

In the previous three sections we presented seven different tag selection methods. Ensemble learning methods that combine the outputs from different “experts” can lead to improved overall performance [9]. Inspired by these methods, we evaluated the predictive power of a simple ensemble method that averages the percentile rankings produced by six of the previous tag selection methods (we call this method **hybrid**)¹¹ We did not include num-apps in the ensemble due to its poor performance.

Table 5 shows a detailed comparison of the precision of all the selection methods we evaluated, including hybrid. Although hybrid yielded lower precision than other methods for the top 256 survey ratings, it out-performed the other methods beyond this threshold. The performance of the hybrid selection method is probably more desirable to system designers: systems will want to show more tags than those associated with the top 256 (0.3%) survey ratings. The performance of this simple hybrid suggests that more sophisticated ensemble learners should be able to provide substantially improved performance.

Now that we have presented our selection methods, we return to research question two:

RQ2: Which tag rating interfaces should designers implement to better select the tags they show to users?

The ratings-based selection methods can only be effective to the extent they have ratings data. Table 4 compares the coverage of each selection method when restricted to only a group’s thumb and survey ratings. The group with both up and down ratings achieves the greatest coverage of the three groups.

In many applications, a method with medium coverage but excellent precision may be more desirable than one with full coverage and low precision. To directly test the effects of different interfaces on selection quality, we constructed a different hybrid method from each of the four experimental groups using only the group’s thumb ratings and used the selection method to predict the group’s survey ratings.

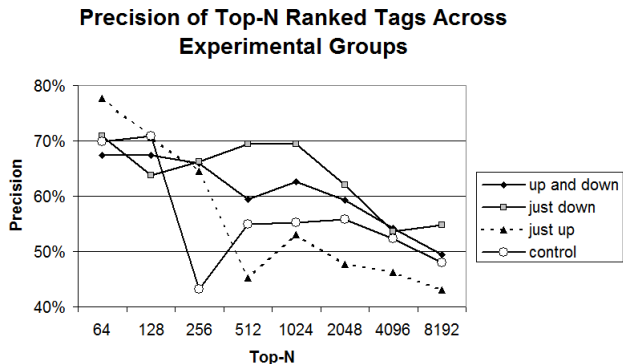
Figure 10 presents the precision results of the hybrid selection method for each of the experimental groups. The high precision of the top 128 ranked ratings for the U group

¹¹Our ensemble method can be viewed as a Bayesian Voting Method [9] in which all ensemble members have similar accuracy and percentile rankings correspond to the probability that a tag is rated positively. Since the probability that a survey rating is rated positively overall is 0.38 (reasonably close to 0.5), this has some empirical justification.

Table 5: Precision of the top- n ranked tags for each tag selection method. The precision was measured over all survey responses. Coverage indicates the percentage of survey responses for which a selection method can generate a prediction. Survey responses rated over three stars were treated as positives.

Method	Coverage	64	128	256	512	1024	2048	4096	8192	16384	32768
num-apps	100.0%	65.5%	65.5%	65.5%	52.9%	35.4%	41.4%	27.1%	36.9%	38.6%	39.1%
num-users	100.0%	65.5%	65.5%	65.5%	65.2%	54.1%	59.5%	56.8%	56.5%	52.5%	47.3%
num-searches	70.2%	87.3%	87.3%	69.6%	67.3%	44.0%	50.3%	45.5%	50.1%	48.6%	44.4%
user-rating	0.5%	81.2%	25.4%	25.4%							
user-avg	1.95%	75.6%	71.9%	69.6%	11.1%	36.7%					
global-avg	92.8%	66.7%	67.3%	67.5%	62.4%	53.6%	54.0%	56.9%	54.5%	53.5%	49.0%
consec-users	92.8%	67.5%	65.3%	65.3%	64.6%	63.8%	59.3%	55.0%	52.6%	48.3%	46.1%
hybrid	100.0%	66.1%	66.1%	66.1%	68.3%	69.5%	66.8%	62.9%	59.0%	57.3%	51.1%

Figure 10: Precision of the hybrid selection method for each of the four experimental groups.



(dotted line with triangle markers) may suggest that thumbs up ratings uncover the best tags, but the poor precision of lower rankings implies that positive ratings do not weed out mediocre and bad tags. The control group without ratings (the solid line with a circle marker) performs well at upper ranks but poorer at middle and lower rankings. The groups with just down ratings (solid line with a square marker) and both up and down ratings (solid line with a diamond marker) both perform quite well at medium and lower rankings.

As stated earlier, precision at lower rankings (e.g. larger n) is more valuable to systems that want to show a large fraction of tags. The down interface, and the interface that used both up and down ratings should be particularly valuable for such systems.

Our results translate into four simple guidelines for designers of tagging systems:

- Systems that support positive ratings should also support negative ratings.** We found that users generate more positive ratings when they could also rate negatively. We also showed that increased rating quantity leads to improved coverage for many tag selection methods. Finally, selection methods using negative ratings, and both positive and negative ratings, performed better than those that just use positive ratings or no ratings at the lower rankings (larger n) critical to real-world systems. This finding is in direct conflict with the policies on many sites that avoid negative ratings for fear that they will drive away users. Our data do not provide tools for directly comparing the benefits of negative ratings for decision-making with the costs of hurting users’ feelings. We believe that most systems should support negative ratings for *objects* such as tags, even if they

do not support negative ratings for *people*.

- Use tag selection methods that normalize each user’s influence.** We found that tag selection methods such as the number of searches or applications per tag are skewed by a small group of “power” users. Tag selection methods that normalize by user, such as the number of users who applied a tag perform better than those that do not.

- Incorporate both behavioral and rating-based tag selection methods.** We found both behavioral and ratings-based tag selection methods to be effective. Table 5 compares precision results across different types of selection methods. Methods based on a user’s own ratings achieved high precision but very low coverage. Methods based on aggregate community behavior and aggregate community ratings performed similarly. Selecting tags based on the number of users who searched for them was particularly precise for those tags ranked highest (87% for the top 128 survey ratings). The hybrid method performs well at the lower rankings (larger n) important for real-world systems.

- Assume that a user’s rating for a particular tag application extends to other applications of the tag.** We found that users generally rate the same tag consistently, regardless of the item it was applied to. For example, 91% of thumb ratings for the same tag, by the same user, but for different items were identical. Although systems may want to allow users to rate individual tag applications, they should interpret a rating for a tag application as strong evidence for a user’s general feeling like for a tag.

We also discovered two surprising characteristics of tag rating in MovieLens. First, users tend to agree more about “bad” tags than “good” tags. We saw evidence for this in tag selection methods (consecutive negative ratings were much more predictive than positive ones), survey results (inter-user agreement was higher among “bad” tags), and general use of the five-star survey scale (there were 7x more one star ratings than two star ratings, but distribution on the high end of the scale was even). Second, in the UD group, although more tag raters rated positively than negatively, a few power users caused the group to generate twice as many negative ratings as positive ones. These results may be specific to communities such as MovieLens that have many low quality tags.

Our research presents several opportunities for future work. Although we focus our analysis of tag selection methods to three basic types of signals (implicit user behavior, a user’s own ratings, aggregate user ratings), more complex techniques may lead to improved accuracy. Our goal was to present system designers with intuitive tag selection meth-

ods that they may easily implement, and to offer both practitioners and researchers insights into several fundamental signals of tag quality. We leave the exploration of more complex algorithms, such as those based on machine learning techniques, as future research.

Second, we would like to validate our techniques using other tagging applications. Surveyers felt that the quality of most MovieLens tags was low enough that they should not be displayed. It would be particularly useful to validate our results in a domain that has a higher ratio of good to bad tags.

Finally, we would like to know whether the design principles we present generalize to other types of community-contributed content such as images, articles and bookmarks. As the size of member-maintained communities grows, communities will require better tools to separate good contributions from bad ones.

9. ACKNOWLEDGMENTS

We would like to thank Dan Frankowski, Shyong Lam, Al Mamunar Rashid, Jilin Chen, and S. Andrew Sheppard for their help in planning the tag rating studies, Sara Drenner for her initial suggestion of rating tags, the rest of Grouplens for their discussion and input, and our MovieLens users for their exuberant tag ratings and survey responses.

This work is funded in part by National Science Foundation, grants IIS 03-24851 and IIS 05-34420.

10. REFERENCES

- [1] A. Arnt and S. Zilberstein. Learning to perform moderation in online forums. *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, pages 637–641, 2003.
- [2] S. Asch. Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs*, 70, 1956.
- [3] M. Balabanovic and Y. Shoham. Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.
- [4] V. Bouthors and O. Dedieu. Pharos, a collaborative infrastructure for web knowledge sharing. In *ECDL*, pages 215–233, London, UK, 1999. Springer-Verlag.
- [5] C. Carson and V. Ogle. Storage and Retrieval of Feature Data for a Very Large Online Image Collection. *Data Engineering Bulletin*, 19(4):19–27, 1996.
- [6] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics in online social communities. In *The European Physical Journal C (accepted)*. Springer-Verlag, 2006.
- [7] D. Cosley, D. Frankowski, S. Kiesler, L. Terveen, and J. Riedl. How oversight improves member-maintained communities. In *Proceedings of CHI 2005*, pages 11–20, New York, NY, 2005. ACM Press.
- [8] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *Proceedings of ACM CHI*, Montreal, CA, 2006.
- [9] T. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15, 2000.
- [10] E. Frank and G. Paynter. Predicting Library of Congress classifications from Library of Congress subject headings. *Journal of the American Society for Information Science and Technology*, 55(3):214–227, 2004.
- [11] N. Glance, D. Arregui, and M. Dardenne. Knowledge pump: Supporting the flow and use of knowledge. In *Information Technology for Knowledge Management*. Springer-Verlag, 1998.
- [12] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science (accepted)*, 2006.
- [13] C. Lampe and P. Resnick. Slash (dot) and Burn: Distributed Moderation in a Large Online Conversation Space. *Proceedings of SIGCHI*, pages 543–550.
- [14] G. MacGregor and E. McCulloch. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library View (accepted)*, 55(5), 2006.
- [15] D. Millen, J. Feinberg, and B. Kerr. Social bookmarking in the enterprise. *ACM Queue*, 3(9):28–35, 2005.
- [16] P. Schmitz. Inducing ontology from flickr tags. *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May*, 2006.
- [17] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *Proceedings of the ACM 2006 Conference on CSCW*, Banff, Alberta, Canada, 2006.
- [18] C. Shirky. Ontology is overrated. http://www.shirky.com/writings/ontology_overrated.html, 2005. Retrieved on May 26, 2007.
- [19] J. Voss. Measuring wikipedia. In *Proceedings of the International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, 2005.

APPENDIX

A. BAYESIAN EXPECTED ENTROPY

Entropy measures the amount of uncertainty associated with a random variable. Entropy is calculated by summing over all possible outcomes $x_1 \dots x_n$:

$$H(X) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (1)$$

In our application, we wish to measure the amount of disagreement in the thumb ratings for a particular tag. Thus, suppose, a tag has 2 positive votes and 3 negative votes. The entropy of the ratings for the tag would be

$$-0.4 \cdot \log(0.4) - 0.6 \cdot \log(0.6) = 0.97 \quad (2)$$

Now suppose that the tag has 20 positive ratings and 30 negative ratings. Since the ratio of positive to negative ratings is the same, the entropy will be the same.

But do we really expect the amount of disagreement to be the same in both cases? In the first example, it is easy to imagine that the “true” underlying ratio of positive to negative ratings is 0.2, 0.5, or 0.7. On the other hand, we have a fair degree of confidence in the entropy measurement for the second example due to its fifty ratings.

A Bayesian approach to entropy calculation treats the up to down ratio itself as a random variable. If we assume that all up/down ratios for tags are equally likely (this is not far from actual reality), then, given u up ratings and d down ratings, the probability of a particular ratio q being f is:

$$p(q = f|u, d) = \frac{f^u(1-f)^d}{\beta(u+1, d+1)} \quad (3)$$

Based on this probability calculation, we can calculate the expected entropy of the ratings by combining a “weighted average” of the entropies for all possible ratios f weighted by the probability of each f as calculated in equation 3:

$$\int_{f=0}^1 p(q = f|u, d) (-f \cdot \log(f) - (1.0 - f) \cdot \log(f)) \quad (4)$$

Using this formulation, we get an expected entropy of 0.84 for the example with five votes and 0.96 for the example with fifty votes, which seems more reasonable.