Tag Expression: Tagging with Feeling

Jesse Vig[1], *Matthew Soukup*[1], *Shilad Sen*[2], *John Riedl*[1]

[1]Grouplens Research
Department of Computer Science and Engineering
University of Minnesota
{jvig,soukup,riedl}@cs.umn.edu

[2]Math, Statistics, and
Computer Science Department
Macalester College
ssen@macalester.edu

ABSTRACT

In this paper we introduce *tag expression*, a novel form of preference elicitation that combines elements from tagging and rating systems. Tag expression enables users to apply *affect* to tags to indicate whether the tag describes a reason they like, dislike, or are neutral about a particular item. We present a user interface for applying affect to tags, as well as a technique for visualizing the overall community's affect. By analyzing 27,773 tag expressions from 553 users entered in a 3-month period, we empirically evaluate our design choices. We also present results of a survey of 97 users that explores users' motivations in tagging and measures user satisfaction with tag expression.

ACM Classification: H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces; H.5.2 [Information Interfaces and Presentation]: User Interfaces

General terms: Design, Experimentation, Human Factors

Keywords: tagging, ratings, user preference, community

1 INTRODUCTION

Tagging sites have flourished across the Web, enabling users to label photos, websites, or even other people with free form descriptors. Most tagging systems are collaborative in nature: tags applied by one user are visible to everyone, the vocabulary of tags evolves as a "folksonomy" [9], and tag cloud visualizations show which tags are most popular for each item [2].

One limitation of existing tagging systems is that they do not explicitly capture user preference. Web users express preference in many ways, such as rating movies on Netflix[1], "digging" articles on Digg[2], or writing book reviews on Amazon[3]. The simplest form of expression is a rating, which may vary in scale from unary (e.g., thumbs-up) to multi-valued (e.g., 5-star) [6]. Text-based reviews and comments, popular on sites such as Amazon and Epinions[4], provide an alternative way to express preferences. Rather than simply indicating how much they like something, users explain why they like or dislike something.

Ratings systems typically accept preferences that are *narrow*, describing only a single dimension, and *explicit*. For example, users of Netflix rate movies on a one-dimensional scale from one to five stars. Because ratings are narrow, producing them requires low cognitive load from the user, but they cannot express the full range of user reaction to an item. Because ratings are explicit, they enable natural summarization in the form of an average rating, and are easily machine-readable for use in automated systems such as recommender algorithms [18]. In contrast, free-form text affords rich expression, but requires more effort from users. Further, because the rating is *implicit* in the text, free-form text does not promote easy summarization in the form of an "average review" and may be difficult for automated systems to interpret.

We introduce a novel interface that enables users to provide ratings across the arbitrary dimensions expressed by the tags applied to an item. Called *tag expression*, this interface bridges the gap between traditional narrow-but-explicit ratings systems and broad-but-implicit text review systems. In this system, users explain their preferences for an item by choosing tags and associating each with one of three affects — like, dislike, or neutral. In this context, affect measures a user's pleasure or displeasure with the item with respect to the tag. For example, someone evaluating the movie "Speed" may express that they like *action*, dislike *Keanu Reeves*, and are neutral about *Sandra Bullock* in this movie. Tag expression also enables users to share their tags and the associated affect with the community. We develop a novel interface for users to apply affect to community tags, and a visualization that presents the aggregate community affect for each tag.

In this paper, we first examine the design space of preference expression and discuss a range of alternatives. We detail our implementation of tag expression and empirically evaluate our design decisions based on a 3-month field study. We present results from a survey exploring users' motivations for using tag expression and their level of satisfaction with it. We also study the impact of tag expression on the overall health of the tagging system. Finally, we draw conclusions for sys-

[1] http://www.netflix.com
[2] http://www.digg.com
[3] http://www.amazon.com
[4] http://www.epinions.com

UIST'10, October 3-6, 2010, New York, NY, USA.
Copyright 2010 ACM 978-1-4503-0271-5/10/10...$10.00.

tem designers based on these findings and discuss potential applications of tag expression.

2 RELATED WORK

2.1 Affect in Interface

Many researchers have explored the role of affect in human-computer interfaces [16]. For example, Breazeal studied ways of developing robots capable of expressing affect to human users. Other researchers have explored the use of machine learning to determine a user's affect automatically. For instance, an educational game may be more effective if it chooses its interventions based on a probabilistic model of the user's current emotional state [5]. Studies have demonstrated that humans learn and interact better when the agents they interact with present affect.

Our work builds on the prior research by exploring how affect changes a tagging interface. In contrast to systems that convey affect or infer affect from user behavior, tag expression enables users to explicitly state their view of the affect of a tag. Some existing systems also support explicit affect expression; for example, LiveJournal[5] allows users to associate a mood with each post they write. Tag expression extends these systems by enabling users to express multiple affects for a single item.

Our definition of affect derives from the three-factor model of emotions developed by Russell et al., which decomposes human emotion into the dimensions of pleasure-displeasure, arousal, and dominance-submissiveness [20]. Breazeal used a similar taxonomy of affect space based on the three dimensions of valence, arousal, and stance [3]. In our work, we consider the single dimension of valence, or pleasure-displeasure, which measures positive-negative emotional state.

2.2 Tag Visualization

Tag clouds have become a common way to show a set of popular tags on a tagging site. Researchers have explored the effectiveness of different tag cloud presentations [19, 21, 2], including changes in font size, weight, color, and the organization of the tags within the cloud. So far there have been a number of interesting alternatives proposed, each with benefits for specific tag navigation tasks, but no clear winner. Because we are not focused on revising the tag cloud model, we use a standard tag cloud with alphabetic ordering and a varying font size that reflects popularity. We integrate affect by varying colors of tags, as described in Section 3.

2.3 Tag Vocabulary

A crucial element of a tagging system is the vocabulary chosen by the users. Studies have shown that users often choose different terms for the same concept [8]. Too restricted a vocabulary will limit the flexibility of the system to appropriately label resources, while too wide a vocabulary will limit the usability of tags for search or filtering. Information theoretic arguments suggest that unmanaged tagging systems may lead to inefficiency, especially over time [4].

The sets of tags that are shown to users in recommendations or drop-down boxes significantly changes the vocabulary the community adopts [9]. The choice of algorithm for presenting tag options to users may influence the steady-state vocabulary used by the community [23]. Our study introduces a new interface for applying tags, which includes an implicit pool of tags the user may choose among. We explore the impact of that interface on the vocabulary used by the community.

3 DESIGN OF TAG EXPRESSION

As a design platform, we used the MovieLens[6] movie recommendation system. MovieLens's primary purpose is movie recommendation: users rate movies on a scale of $1/2$ to 5 stars and receive recommendations in return. MovieLens was launched in 1997, and it attracts approximately 3200 active users per month. Since tagging was first introduced in January 2006, 4,745 users have created 174,240 tag applications resulting in 17,991 distinct tags (a *tag* is a distinct word or phrase, while a *tag application* is a user's association of a tag to a movie represented by a (*user*, *tag*, *movie*) triple.).

MovieLens users may apply tags to movies on two different screens. On the *movie details* page, users see detailed information about a single movie such as genre, cast, director, as well as a list of the top 20 tags.[7] On the *search results* page, users see high-level information for the movies returned from a particular search, including the top 3 tags for each movie. We focus our design discussion on the movie details page, where 86.2% of tagging activity occurred under tag expression. Under the previous tagging system, users created tag applications on the movie details screen by typing in an auto-complete text box or simply clicking on an icon next to each existing tag. More details about the MovieLens system and its tagging implementation are presented by Sen et al. [23].

Our design of tag expression focused on three elements: preference dimensions, affect expression, and display of community affect. In the following sections we define each design element, outline the alternative we considered, and explain our design decisions. Later in Section 5, we evaluate each of these decisions based on user activity and results from the survey. For several design elements we refer to the tag expression component of the movie details page shown in Figure 1.

3.1 Preference dimensions

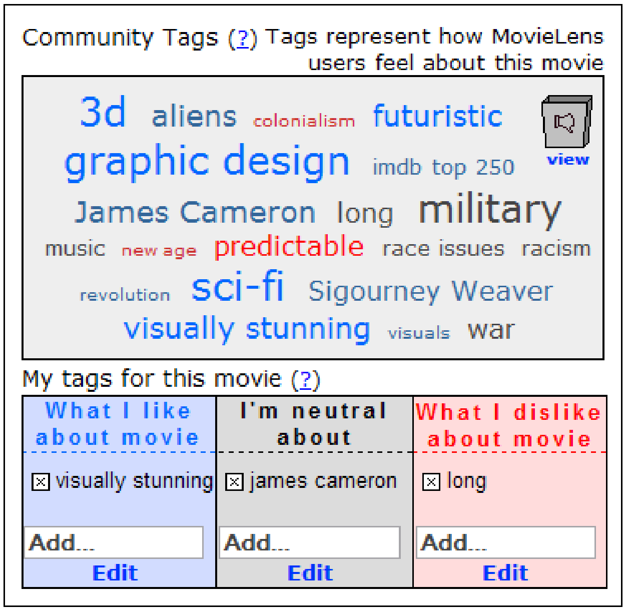
Before settling on using tags to express preferences, we considered other entities through which users might express preferences. We refer to these entities as *preference dimensions*. Rating systems typically capture a single preference dimension that describes the user's overall sentiment towards the item. For example, a user on MovieLens might rate the movie "The Usual Suspects" with 4 stars. With a broader set of preference dimensions, a user could rate the *plot* of "The Usual Suspects" as 5 stars, but the *action* as only 3 stars.

The question is how to choose the appropriate set of preference dimensions. With an expert-based approach, domain experts or system designers hand pick the dimensions; for example, designers of a movie website might include the dimensions of genre, plot, acting, level of violence, etc. Ex-

[5] http://www.livejournal.com

[6] http://www.movielens.org

[7] Top tags selected based on a tag quality metric developed by Sen et al. [24]



Community Tags (?) Tags represent how MovieLens users feel about this movie

3d aliens colonialism futuristic graphic design imdb top 250 James Cameron long military music new age predictable race issues racism revolution sci-fi Sigourney Weaver visually stunning visuals war

view

My tags for this movie (?)

What I like about movie	I'm neutral about	What I dislike about movie
☒ visually stunning	☒ james cameron	☒ long
Add...	Add...	Add...
Edit	Edit	Edit

Figure 1: Tag expression component on the detailed movie information page for *Avatar*. Tags that describe an aspect of the movie liked by the MovieLens community are colored blue (*3d*), neutral aspects are colored gray (*military*), and negative aspects are colored red (*predictable*). A user can add their own affect (*visually stunning*, *james cameron*, *long*) by dragging a tag to one of the three affect containers, or by typing a new tag in an affect container's text input box.

perts are likely to choose a compact and consistent set of preference dimensions, and less effort is required from users. However, experts cost money, and they may not be able to anticipate the varied and changing interests of the user community.

In a user-based approach, the users themselves define the space of preferences. Tagging is a common approach for representing a set of user-defined dimensions. Although tagging is generally used to support information retrieval, it is also well-suited for articulating preferences, for several reasons. First, tags are atomic: unlike other text-based forms of expression such as user reviews, each tag typically represents a single concept, enabling users to express preference along each dimension independently. Second, tagging supports collaboration between users. A tag applied by one user may be reused by another, lowering the cognitive load of preference expression, and tags provide common units of expression around which the overall community's preferences may be aggregated and displayed. However, tagging systems face many challenges, including redundant tags (*dark comedy*, *black comedy*), low quality tags (*bah*), and personal tags (*Erlend's DVDs*).

We opted for a user-centric approach that employs tags as preference dimensions, for several reasons. First, preferences are inherently personal and tags provide users the flexibility to express their diverse interests. Second, the community-oriented features of tagging can enrich preference expression by making it more social. Third, the tremendous popularity of tagging systems invites exploration of broader applications of tagging beyond information retrieval.

3.2 Expressing affect

A second design decision is how users associate affect, or pleasure-displeasure, with each of the preference dimensions. Here we consider three aspects of affect expression: rating scale, granularity, and interface.

Rating scale. The rating scale describes the range of values that a user may associate with a preference dimension, i.e. tag. We considered a variety of rating scales for affect, but focused on the most commonly-used scales: unary (thumbs-up), binary (thumbs-up / thumbs-down), and many-valued (five-star) [6]. We felt that a five-star scale would be too complex — in terms of both interface design and cognitive complexity. We preferred a binary scale over a unary scale based on Sen et al.'s finding that users rate significantly more tags with a binary scale [22]. Since tag expression entirely replaced the existing tag creation mechanism, we wanted a way for a user to create a tag with neither positive nor negative affect. Therefore, we decided to use a ternary scale with positive, negative and neutral options.

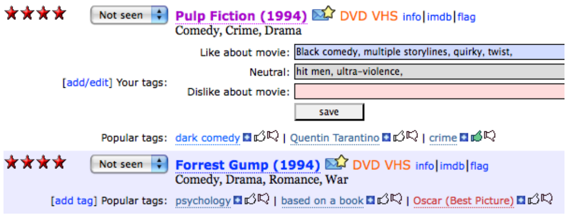
Granularity. A second question concerns the level of granularity, or specificity, at which users express affect. With a *per-tag* approach, a user associates affect with a tag in isolation, independent of any particular item. For example, a user may express that they simply like or dislike *Arnold Schwarzenegger*. With a *per-item-tag* approach, users associate affect with a (*movie*, *tag*) pair. In this scheme, a user might express that they like *Arnold Schwarzenegger* for the movie "Terminator" but dislike *Arnold Schwarzenegger* for the movie "Kindergarten Cop".

We opted for the per-item-tag approach, for two reasons. First, per-item-tag provides greater flexibility of preference expression, as illustrated by the above example. Second, using the least constrained model allowed us to observe the specificity of user preference in practice. In Section 5, we analyze whether individual users tended to choose uniform affect for the same tag or instead chose affect based on the movie.

Interface. We considered two designs for the affect expression interface. In the first option, MovieLens would display a rating widget alongside each of a movie's tags. In the second mechanism, MovieLens displays an *affect container* for each of the three affect values, and asks users to create affects in a particular container. We selected the latter design because it required less screen space, allowing us to specify descriptive labels for each of the affects: "What I like about movie", "I'm neutral about", and "What I dislike about movie." To reduce the effort of creating tags, MovieLens allows a user to drag tags applied by other users to an affect container. One-click tag reuse had also been available in the previous tagging system through an add button next to each tag. Users may also create brand new tags with a particular affect by typing them in the auto-complete text input box corresponding to the affect. The three affect containers can be seen at the bottom of Figure 1.

Aggregation	Result
Plurality vote	*negative*
Histogram	(*positive*: 3, *neutral*: 1, *negative*: 4)
Mean value	-0.125

Table 1: Results of each affect aggregation function, for the tag *violent* on the movie “In Bruges”. 3 MovieLens users expressed positive affect for this tag, 1 user expressed neutral affect, and 4 users expressed negative affect.



Pulp Fiction (1994)
Comedy, Crime, Drama
[add/edit] Your tags: Like about movie: Black comedy, multiple storylines, quirky, twist,
Neutral: hit men, ultra-violence,
Dislike about movie:
Popular tags: dark comedy | Quentin Tarantino | crime

Forrest Gump (1994)
Comedy, Drama, Romance, War
[add tag] Popular tags: psychology | based on a book | Oscar (Best Picture)

Figure 2: Tag expression component for the search results page.

3.3 Displaying community affect

One of the goals of tag expression is to display an aggregate representation of the community’s preference for a movie. Here we consider methods for aggregating and visualizing community preference.

Aggregation function. The aggregation function summarizes the affect values applied to a (*movie*, *tag*) pair into a single value or a small set of values. Motivated by the use of voting theory in information aggregation [7] and preference aggregation [13], we considered several voting-based approaches for aggregating affect. *Plurality voting* is a single-winner voting system that selects the most popular choice. Using this approach, the affect aggregation function would simply choose the affect value applied by the most users to the (*movie*, *tag*) pair. *Proportional representation* is a voting system in which alternatives are represented in proportion to their popularity. Traditionally, this is a multi-winner voting system in which each discrete alternative is selected in quantity proportional to the number of votes it received. Following this model, the tag expression system might display a histogram showing the distinct affect values applied and their relative frequencies. We also consider a variant of proportional representation that relaxes the assumption of discrete, indivisible alternatives by mapping each affect value to a real number (positive: +1, neutral: 0, negative: -1) and taking the mean of the results. Figure 1 illustrates each of these aggregation methods.

One interesting question is how well each aggregation method resolves the tension between accurately reflecting the overall preference of the community for a tag, and showing the full breadth of community opinion about the tag. Plurality vote summarizes overall preference, but gives no indication of the range of opinion. Histogram shows the range of opinion, but doesn’t compute an overall community preference. Mean value achieves both goals to varying degrees; it computes a single value for overall preference that is equally influenced by all preferences expressed. However, the mean value conveys less information about the breadth of opinion than a histogram, as illustrated by the example in Table 1. In that example, the mean value (-0.125) is the same as if 7 users had applied neutral affect and one user had applied negative affect. In contrast, the histogram reveals that this is a highly controversial tag where user opinion is roughly split between positive and negative affect.

We chose to use mean value, for several reasons. First, it balances the goals of capturing overall preference as well as breadth of opinion. Second, it presents a simple conceptual model to users since there is a single value associated with each tag, as opposed to a histogram, which shows up to three values for each tag. Since many tags are displayed on the screen at once, users may be overwhelmed by processing multiple values for each tag. Third, it is consistent with common approaches for estimating community preference for an item, where the aggregation function takes the mean of all ratings. To account for the uncertainty arising from small sample sizes, we use Laplacian smoothing by adding one positive, one neutral, and one negative affect to the values being averaged.

Visualization. Two common methods for visualizing tags applied to an item are tag clouds and tag lists. An advantage of tag clouds is that they can convey tag popularity through font size while maintaining an alphabetic order for retrieval of specific terms. We chose to use a tag cloud in our implementation of tag expression, but system designers may also consider tag lists sorted alphabetically or by popularity.

We needed to augment the tag cloud with information that summarizes the affect applied to these tags. We considered several design options for visualizing aggregate affect, including a star-based system, manipulation of a tag’s font size, and manipulation of a tag’s color. We chose color to convey aggregate affect because it requires less space than the star-based system, and because font size already commonly conveys tag frequency in tag clouds. For colors, we initially considered red for negative and green for positive based on their common use in other applications (e.g. stop lights in the United States). However, since 7% of male adults are color blind and cannot distinguish between red and green [14] we replaced green with blue, the color that contrasted second most strongly with red. We used gray for neutral affects. We employed the following mapping from aggregate affect to color: -1.0 to -0.6: red, -0.6 to -0.2: dull red, -0.2 to 0.2: gray, 0.2 to 0.6: dull blue, 0.6 to 1.0: blue. In addition to conveying mean value of affect, we also considered visualizations for depicting affect variance. For example, highly controversial tags with a high variance of affect values for a particular movie might be shown with a red-and-blue checked background. We opted not to include such a visualization because we felt it would make the display too complex when combined with varying font size and color.

3.4 Other considerations

As discussed earlier, we focused our design on the movie detail page, where 86% of tagging occurred. For consistency we also incorporated tag expression into the search results page, shown in Figure 2. Since the search results screen displays information about many movies, screen space is limited. We display three tags for each movie[8], each colored based on aggregate community affect. Users can also add new tags with a particular affect using the auto-complete text input box associated with each of the affects.

We created two additional temporary pages to promote tag expression. The *introduction page* was shown once to all users who logged in after we launched tag expression. The page described the new feature and it also stated that tag expression data might someday be used to improve users' movie recommendations. The *affect migration* page enabled a user to add affect to tags they had applied before tag expression was launched.

4 EXPERIMENTAL METHODS

We conducted a field study of tag expression on the MovieLens website, in which we empirically evaluated tag expression based on activity logs and survey data. The primary data source for our analyses comprised activity logs collecting during a 3-month period that tag expression was in place, running from May 27, 2009 to August 27, 2009. These logs track the tags and associated affect that users applied during this time. Although we introduced tag expression in April 27, 2009, we excluded the first month of activity in order to control for the spike of activity that typically comes with a new interface.

We also compare tagging activity under tag expression to tagging activity under the previous tagging system. For this comparative analysis, we included activity logs from February 2006 through April 2009, the range of time that the previous tagging system was in place (excluding the first month). A longitudinal study such as this presents many challenges. First, the data collected under the previous tagging system spans three years, during which time users' tagging behavior and perceptions of tagging systems may have changed. We found that tagging activity on MovieLens was fairly consistent over time; users created 54,867 tag applications in the first 19 months that the previous tagging system was in place, compared 57,612 in the last 19 months. Overall user activity also stayed fairly constant, with 3,184 users logging in per month under the previous tagging system, compared to 3,201 users per month under tag expression.

Another challenge of comparing the two systems is that tag expression brought other interface changes besides the primary change of introducing affect into tagging. First, the representation of tags changed from a simple list to a tag cloud with varying font sizes and colors. Not only does this change the look of tags, it also increases the size of the tagging area. Second, the introduction of affect containers changed tagging from a labeling task to a classification task (*like*, *neutral*, *dislike*). Some of the changes in tagging behavior may have occurred in any interface where users classified tags into multiple categories, for example *factual*, *subjective*, *personal*. The results of our comparative study should be interpreted holistically in respect to the entire set of interface changes.

In addition to analyzing tagging behavior, we conducted an online survey to measure user satisfaction and explore motivations for using tag expression. In the survey, subjects first compared the tag expression interface to the previous tagging interface, labeled as Option A and Option B respectively. Subjects compared the interfaces based on overall preference and in respect to specific tasks (self-expression, deciding about movies, learning about movies, finding movies), selecting one of the following responses: *A is much better*, *A is better*, *Both are about the same*, *B is better*, *B is much better*. Subjects also responded to two statements that targeted features unique to tag expression: (1) *I like seeing the colors that show how other MovieLens members feel about the tags* and (2) *I like the ability to influence the color of the tags shown to the community*. Finally, subjects rated several reasons for using tag expression (contribute to the community, self-expression, improve recommendations, organize movies, fun, curiosity) using a 5-point Likert scale (*strongly agree*, *agree*, *neutral*, *disagree*, *strongly disagree*).

Reason	% Agree
Better Recommendations	**78.5%**
Contribute to Community	**75.5%**
Curiosity	**64.5%**
Self-expression	**60.2%**
Fun	**48.9%**
Organize My Movies	**27.8%**

Table 2: Percentage of subjects who agreed with reason for tagging (N=97). Differences in agreement greater than 12% are statistically significant (p<0.05), based on the Z-test of two proportions.

5 EMPIRICAL EVALUATION

In this section we evaluate our design decisions empirically, based on activity logs recorded during the 3-month field study and the results of the user survey. We focus on the three design elements discussed earlier: preference dimensions, affect expression, and display of community affect.

5.1 Preference dimensions

We explored whether users actually used tags to express preferences. We found that users expressed positive or negative affect for the majority (61.8%) of tag applications, suggesting that preference expression had been successfully adopted into the tagging task. Further, survey results shown in Table 2 suggest that users largely tagged *in order to* express preference, specifically for improving recommendations (78.5% of subjects) and self-expression (60.2% of subjects). Contributing to the community (75.5%) was also a primary motivation for using tag expression, consistent with other work showing the importance of social motivators in tagging [1]. Some of the desire to contribute may also relate to preference expression: 63.9% of survey participants liked the ability to influence the color of the tags shown to the community.

5.2 Expressing affect

Affect distribution. As discussed earlier, we chose to represent affect on a ternary scale (*positive, neutral, negative*). One indication of a good scale is that users take advantage of the full range of values. Under tag expression, users created 53.4% of tag applications with positive affect, 38.2% with neutral affect, and 8.4% with negative affect. One im-

[8] Based on a tag quality metric developed by Sen et al. [24]

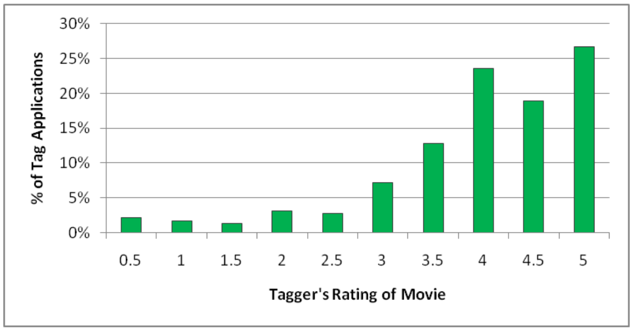


Figure 3: Distribution of tag applications over ratings.

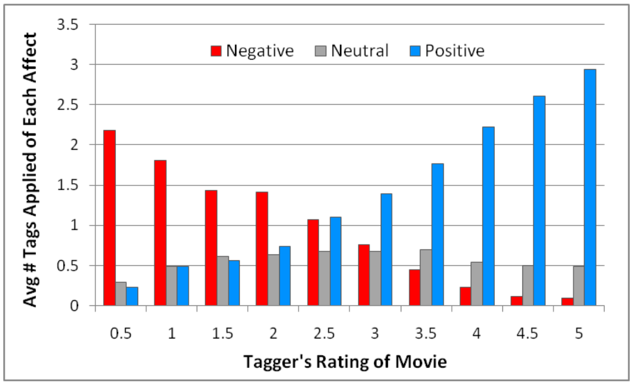


Figure 4: Average number of tags of each affect a tagger applies to a movie, as a function of the tagger's rating of the movie.

plication is that the neutral option was a good design choice, given that over a third of tags were applied with neutral affect. The low proportion of negative affect is largely due to which items users tagged. As Figure 3 shows, users applied over twice as many tags to movies they rated 4 stars or higher compared to movies they rated less than 4 stars.[9] Users' movie preferences strongly influenced their choice of affect, as demonstrated by Figure 4. The number of tags applied with positive (negative) affect increases (decreases) with movie rating in a nearly linear fashion ($R^2 = 0.97$ in both cases). Thus although users apply negative affect infrequently overall, they rely on it heavily for movies toward the lower end of the rating spectrum.

We wondered whether certain tags elicited a wider range of affect values than others. To identify the tags with greatest affect variation, we grouped affects applied to a given tag and measured the entropy of the affect distribution. We used the Bayesian expected entropy measure proposed by Sen et al., which measures Shannon entropy but compensates for uncertainty resulting from limited sample sizes [22]. The tags with greatest affect variation were *bleak* (entropy 0.982), *acting* (0.980), *gore* (0.974), *keanu reeves* (0.972), *boxing* (0.970), *julia roberts* (0.969), *kevin costner* (0.968), *eddie murphy* (0.968), *incest* (0.966), and *torture* (0.964). Each tag appears to fall into one of the following categories: (1) tags for which users have differing preferences (*bleak*, *gore*, *boxing*), (2) neutral tags that typically require a qualifier (*acting*), (3) tags for specific actors (*keanu reeves*, *julia roberts*, *kevin costner*, *eddie murphy*), or (4) tags representing culturally sensitive topics (*incest*, *torture*). One possible explanation for the last case is that users may sometimes express affect toward the subject itself ("I dislike torture") and other times toward the handling of the subject ("I like how this film addresses the subject of torture").

The relative benefits of negative versus positive affect depend on how tag expression is used. Systems that use tag expression to generate a top-N list of recommendations may find positive affect more informative because it reveals candidates for the top-N list. Systems that predict a user's preference for an arbitrary item – such as a search result – may benefit from understanding both a user's likes (positive affect) and dislikes (negative affect). System designers may choose design elements that encourage the type of affect most beneficial to their system. For example, designers who want a higher proportion of negative affect might prompt users to tag items that they rate lower than 3 stars.

Figure 4 suggests that users choose affect as a way of explaining how they felt about a movie: users who liked a movie overwhelmingly expressed positive affect for its tags, while users who disliked a movie overwhelmingly expressed negative affect. Prior research has shown that tags can be a powerful way to explain to a user why a movie was recommended [26]. Users seemed to be intuitively developing such explanations for themselves. Understanding how users construct these explanations may provide insight for designers of recommender systems that provide explanations to users. For example, one issue in designing explanations is how to choose the proper mix of positive and negative aspects to show for an item with a certain predicted rating. One approach would be to use the proportions of positive and negative affect in Figure 4 corresponding to the predicted rating.

Granularity of affect. As discussed earlier, we chose a per-item-tag model, where users associate affect with a (*movie*, *tag*) pair rather than a tag in isolation. To evaluate this design choice, we investigated how users chose affect when applying the same tag to different movies. We found that in 75.9% of cases, users chose uniform affect when applying a tag to multiple movies. These data suggest that the per-tag model may have been sufficiently expressive to capture user preference. However, the per-item-tag was needed for a non-trivial percentage of cases (24.1%). System designers seeking the benefits of both approaches might consider a hybrid system where users may express preferences for tags in isolation or in the context of a specific item.

5.3 Displaying affect

We chose mean value as an aggregation function because it balances the goals of showing overall community preference on the one hand and representing the breadth of opinion on the other. In this section we evaluate how well tag expression accomplishes each of these goals.

Breadth of opinion. We investigated how well the displayed affect reflected the breadth of opinion of individual taggers. To get a sense of the diversity of displayed affect, we counted

[9] For this analysis we excluded the 23% of tag applications where the tagger had not rated the movie.

the number of distinct (*movie*, *tag*) pairs of each color across all movies.[10] We found that 48.4% displayed positive affect (blue or dull blue), 44.3% showed neutral affect (gray), and 7.3% showed negative affect (red or dull red). Figure 5 shows the distribution of displayed affect as a function of a movie's average rating. Overall, negative affect is by far the least commonly displayed affect. Even for movies with an average rating of 2 stars, positive affect exceeds negative affect.

The scarcity of negative affect in the display does not necessarily imply that the system is failing to represent the full range of opinion that users express. Rather, it may simply reflect the fact that users overwhelmingly express positive affect (see Section 5.2). To better understand the relationship between the affect that individual users express and the aggregate affect that the system displays, we computed the relative proportion of displayed affect as a function of users' expressed affect, as shown in Figure 6. When a user applies positive affect to a tag, there is a 94.3% chance that the aggregate affect subsequently displayed will be positive (blue or dull blue). However, after a user expresses negative affect, there is only a 69.1% chance the aggregate affect will appear as negative (red or dull red). The difference in percentages are statistically significant ($p < 0.001$, Z-test of proportions). Thus the aggregation mechanism may be disproportionately suppressing negative affect from the display. Seeing more positive affect may in turn influence users to apply more positive affect. This type of cascade effect has been reported in both tagging and rating systems [23, 6].

System designers who wish to better represent minority opinions may prefer to use a histogram showing all affects applied and their respective frequencies, as discussed in Section 3. Alternatively, designers might consider a personalized display of affect that assigns higher weight to affect applied by other users with similar tastes. This is analogous to user-based nearest-neighbor recommender algorithms, which predict ratings based on a weighted average of ratings from similar users [17]. With a personalized display of affect, minority opinions will be better represented to users likely to share those opinions.

Overall community preference. We also explored how accurately the aggregation function represents the community's *overall* preferences. As discussed in Section 3, we chose an aggregation function that computes the mean affect expressed by users for a (*movie*, *tag*) pair. While this aggregation function summarizes the affect values that users express, the users who choose to express affect constitute only a fraction of the overall user community. If these taggers' preferences do not reflect the overall community's preferences, then the aggregate affect will not accurately represent the overall community's preferences either.

To measure how well taggers' preferences reflect the community's preferences, we computed the average rating of each movie among just its taggers as well as over the entire user population.[11] Averaged across all movies, the mean movie rating among taggers was 3.86 compared to 3.63 among all users ($p < 0.001$, two-tailed t-test); the difference between the means is equivalent to a 25% difference in percentile in average movie rating. Researchers have found users are more likely to rate items they like [12]; our results suggest that a positivity bias is also present among taggers. Given the relationship between movie rating and affect shown in Figure 4, taggers are therefore expressing more positive and less negative affect than they would if they accurately represented the overall community's preferences.

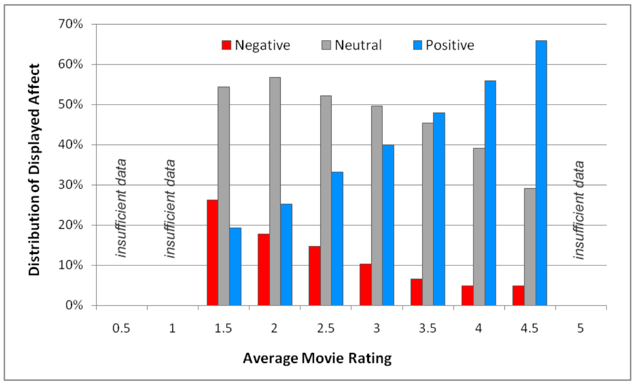


Figure 5: Distribution of community affect for a movie as a function of average movie rating, rounded to nearest one-half star. Shows distribution for rating levels that have at least 10 associated movies.

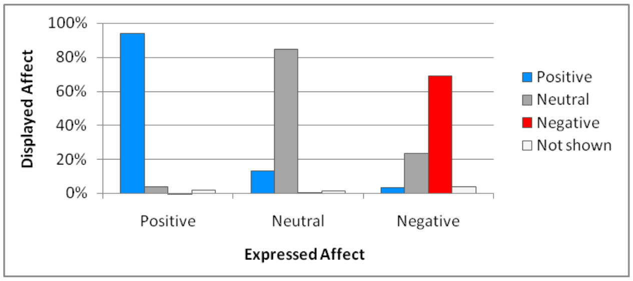


Figure 6: Relationship between affect expressed and final aggregate affect displayed.

To better understand the positivity bias in tagging, we estimated the conditional probability of a user tagging a movie given their rating of the movie. For each user who tagged under tag expression, we grouped the movies they rated by rating (½ to 5 stars) and computed the proportion of each set of movies that the user also tagged.[12] We then averaged the results over all of the taggers. Figure 7 shows the results. If no selection bias existed, the probabilities would all be equal. The estimated tagging probability is approximately quadratic ($R^2 = 0.93$) in rating and attains its maximum at the high end of the rating scale and its minimum towards the middle of the rating scale.

Tagging system designers who wish to more accurately represent overall community preference need to account for this selection bias. One approach is to weight tag applications in

[10] We included the 20 most popular tags for each movie because tagging systems typically filter tags based on popularity.

[11] We included all movies tagged by at least 3 users and with at least 100 ratings.

[12] We only included users at a rating level if they rated at least 2 movies at that level.

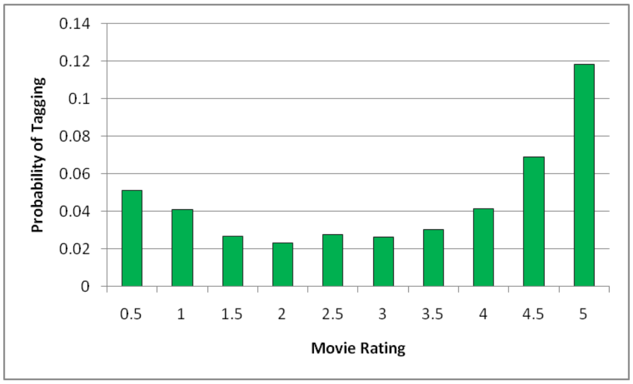


Figure 7: Probability of a user tagging a movie given their rating of the movie. Difference from uniform is statistically significant ($p < 0.001$, one-way ANOVA)

	Prefer A	Neutral	Prefer B
Overall preference	79.2%	8.3%	12.5%
Self-expression	85.5%	9.3%	5.3%
Decide about movies	59.8%	32.0%	8.3%
Learn about movies	59.8%	26.8%	13.4%
Find movies	40.6%	47.9%	11.5%

Table 3: Percentage of subjects who prefer Option A (tag expression) versus Option B (previous tagging system).

the aggregation process. For example, tag applications could be weighted so that the number of tag applications from users who rated a movie, say, 3 stars is proportional to the overall number of users who rated the movie 3 stars. Alternatively, systems designers who have in mind an ideal distribution of positive, negative, and neutral affect could simply weight each affect value by a constant factor to achieve this distribution.

5.4 User satisfaction.

Table 3 shows survey results measuring user satisfaction with tag expression compared to the previous tagging system. Subjects preferred tag expression overall by a substantial margin (79% vs. 13%). Most notably, 86% preferred tag expression for self-expression, compared to just 5% who preferred the old tagging system. This result is consistent with the different tagging models used by the two systems: while users could express themselves in the old tagging system by using subjective tags such as *witty* or *boring*, tag expression is designed explicitly for expression. Depending on the choice of affect, even factual tags such as *Tom Hanks* or *romance* may provide an outlet for users to express themselves. One subject wrote *"I like option A because it allows me to mark tags as like, dislike, and neutral."*

Subjects also preferred tag expression for learning about movies and deciding whether to watch them. According to one subject, *"Option A is far better than Option B; it shows more information and allows much clearer ideas of what the community thinks about a film."* The use of color to represent tag affect appears to play a key role in this: 76% of subjects liked seeing the colors that showed how others felt about movies. Subjects preferred tag expression for finding movies, but by a much narrower margin; though only 12% preferred Option B, nearly 50% of users were neutral about the choice of interface for the finding task.

6 IMPACT ON TAGGING SYSTEM

Tag expression repurposes tagging for the task of preference expression. But tagging systems must continue to support traditional tagging tasks such as navigation and organization [11, 15]. Therefore an important question is how tag expression impacts the overall health of the tagging system. In this section we study the impact of tag expression on the volume, diversity, quality, and types of tags users apply. The results we present should be interpreted in the context of the limitations of the field study discussed in Section 4.

6.1 Tagging volume

We found that users applied 9,273 tags per month under tag expression, compared to 3,031 per month under the old system, a 206% increase. To understand the nature of the increase in tagging volume, we decomposed each month's tagging activity into three factors: (1) the number of users who tagged, (2) the number of movies tagged per tagger, and (3) the number of tags applied per tagger per movie. We averaged the three factors over each of the two time periods, and found that significantly more users (+44%) tagged under tag expression and that they applied significantly more tags (+68%) to each movie they tagged ($p < 0.05$ and $p < 0.001$, two-tailed t-test).

We also explored the role of tag reuse in the increased rate of tagging. A user *reuses* a tag when they apply the tag to an item that already has the tag. Prior to tag expression, 26.7% of tag applications came from reused tags. Under tag expression, the reuse percentage jumped to 69.2%. This difference is statistically significant ($p < 0.001$, Z-test of two proportions). In both tagging systems, reuse was available with a single click. One possible explanation for this increase is that the introduction of affect gives users a new reason to reuse existing tags, because they are attaching their personal opinion to the tag. Some tag reuse is valuable because it reveals the most popular tags and encourages vocabulary convergence. However, a potential risk of increased tag reuse is that users may stop applying tags that are new to items. We found this not to be the case; applications of tags new to movies rose from 2,220 per month under the old tagging system to 2,869 per month under tag expression (difference not statistically significant). So, even though most of the tag applications represent reuse, users created sufficiently more tag applications overall that more original tags were added each month than with the original system.

6.2 Tag diversity

A related question is how tag expression impacts the overall diversity of the tagging vocabulary. We measured the diversity of tag applications created with tag expression based on (1) number of distinct tags and (2) Shannon entropy of tag applications over distinct tags. We calculated both metrics for each month of tagging activity under tag expression and averaged the results. We performed the same computation for tag applications created under the previous tagging system. Both diversity measures were significantly higher under tag expression. The number of distinct tags applied per

	Before	After	Change
Factual	80.2%	87.2%	+8.7%
Subjective	14.6%	12.1%	-17.1%
Personal	5.3%	0.8%	-84.9%

Table 4: Distribution over tag classes before and after launch of tag expression. Each difference was statistically significant (*p*<0.001) based on Z-test of two proportions.

month rose by 119.5% ($p < 0.001$, two-tailed t-test), while entropy rose by 12.4% ($p < 0.01$, two-tailed t-test). The increased volume of tagging may be at least partially responsible for the increase in diversity. Note that it is not clear that a more diverse set of tags is always better. A more focused vocabulary may lead to more effective searches under some conditions. Future research should explore the tradeoffs between focus and diversity.

6.3 Tag quality

We wished to explore whether the individual tags applied under tag expression were more or less useful than those applied through the old system. Since a primary function of tagging systems is to help users search for items, we use tag "searchability" as a measure of tag quality. We define the searchability of a *tag* as the number of distinct users who have searched for items using that tag, based on tag searches conducted on MovieLens between December 2007 and May 2009. We define the searchability of a *set of tag applications* as the average searchability of the tag associated with each tag application. We found that tag applications created during tag expression had 48% higher mean searchability ($p < 0.001$, two-tailed t-test) and 168% higher median searchability than those created under the old tagging system. The improvement in tag quality may be partially due to the increases in tagging volume and diversity: with more tags to choose from, users may make better selections when reusing existing tags.

6.4 Types of tags

In earlier work on tagging vocabulary, Sen et al. divided tags into three broad classes[13]: factual (e.g. *drama, james bond, crime*), subjective (e.g. *funny, overrated, must see!*), and personal (e.g. *in netflix queue, mydvds, get*) [23]. According to their study, users generally preferred factual tags to subjective ones and subjective tags to personal ones. One potential risk of tag expression is that users may apply more subjective tags compared to a traditional tagging system because the subjective tags reflect their personal preferences better than the factual tags. Increasing the fraction of subjective tags might make the vocabulary worse for most users.

Table 4 shows the proportion of tags of each type applied before and after the launch of tag expression. The biggest percentage change (-84.9%) occurred for personal tags, possibly due to the low proportion of users tagging to organize movies as shown in Table 2. Since users liked personal tags least, this represents an improvement to the tagging vocabulary. Surprisingly, the proportion of subjective tags declined after the launch of tag expression. One explanation is that in tag expression users can express their subjective opinions by applying affect to a factual tag. While only 12.1% of tag applications had a subjective *tag*, 62% included a subjective (positive or negative) *affect*. By enabling users to express affect separate from the tag itself, tag expression provides an outlet for expression that does not compromise the tagging vocabulary.

7 CONCLUSIONS

Tag expression is a novel user interface tool that enables users to share the way specific tags express their positive or negative feelings about specific movies. Because these preferences are shared, users can see in aggregate how the overall community feels about a movie. During three months of usage, tag expression has been very popular with MovieLens users, tripling monthly tagging activity and increasing the number of active taggers per month by 44%. Further, 79% of users surveyed preferred tag expression to a traditional tagging system.

We found that tag expression encouraged more community-oriented tagging behavior. Taggers reused others' tags more frequently compared to the previous tagging system, and taggers applied tags that were more "searchable" by other users. Further, users of tag expression applied proportionally *fewer* personal tags than users of the previous tagging system. Users shared not only tags but also opinions, with 62% of tag applications expressing positive or negative affect. Overall, users valued the exchange of opinions: 76% liked seeing the colors that showed how others felt, and 64% liked being able to influence the colors showed to others. Tagging system designers who wish to emphasize the community aspects of their systems may find tag expression a rich way to encourage more community-directed tagging behavior.

Future work may explore recommender algorithms that utilize tag expression data. The level of detail provided by tag expression may help improve recommendation accuracy, particularly in domains such as housing or travel where users may rate only a small number of items. While previous researchers have investigated using tags in recommender algorithms [25, 27], additional work is needed to develop recommender algorithms that consider the affect of a tag. Further, recommender systems might use data collected from tag expressions to generate more detailed predictions. For example, a movie recommender might predict that a particular user will like the *drama* of "Saving Private Ryan", but dislike the *violence*. More fine-grained predictions would help users choose items based on mood, and also serve to explain the recommendation to users [26, 10].

Tag expression has a wide variety of potential applications. News websites could use tag expression to collect and visualize public opinion on the day's events. Aggregator sites such as Digg[14] might use tag expression as a way to aggregate interest along multiple dimensions. Rather than selecting just the most highly rated content, aggregator sites could highlight the most *intelligent*, *witty*, or *uplifting* content, based on

[13] These codings covered 36% of tag applications before the launch of tag expression, and 58% of tag applications after

[14] http://www.digg.com

the number of times the respective tag was applied with positive affect. Social bookmarking sites could use affect as an additional dimension by which to index and retrieve items. For example, a user of CiteULike[15] may wish to find examples of papers with a strong introduction; with the added dimension of affect, they could retrieve all papers where *introduction* is associated with positive affect. We encourage system designers to explore these and other alternatives.

8 Acknowledgments

The authors thank the members of Grouplens for their feedback and assistance. This paper is funded in part by National Science Foundation grants IIS 03-24851, IIS 05-34420, IIS 09-64695, and IIS 09-64697.

9 REFERENCES

1. M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980. ACM, 2007.
2. S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: The effect of visual features on tag cloud selections. In *HT '08: Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, pages 193–202. ACM, 2008.
3. C. L. Breazeal. *Sociable machines: Expressive social exchange between humans and robots*. PhD thesis, 2000.
4. E. H. Chi and T. Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *HT '08: Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, pages 81–88. ACM, 2008.
5. C. Conati. Probabilistic assessment of user's emotions in educational games. *Applied Artifical Intelligence*, 16(7&8):555–575, August 2002.
6. D. Cosley, S. K. Lam, I. Albert, J. Konstan, and J. Riedl. Is seeing believing? How recommender system interfaces affect users' opinions. In *CHI*, 2003.
7. E. David and K. Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
8. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, 1987.
9. S. Golder and B. A. Huberman. The structure of collaborative tagging systems, Aug 2005.
10. S. J. Green, P. Lamere, J. Alexander, F. Maillet, S. Kirk, J. Holt, J. Bourque, and X.-W. Mak. Generating transparent, steerable recommendations from textual descriptions of items. In *RecSys '09: Proceedings of the third ACM conference on Recommender systems*, pages 281–284. ACM, 2009.
11. T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools: A general review. *D-Lib Magazine*, 11(4), April 2005.
12. B. M. Marlin and R. S. Zemel. Collaborative filtering and the missing at random assumption. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.
13. J. Masthoff. Group modeling: Selecting a sequence of television items to suit a group of viewers. *User Modeling and User-Adapted Interaction*, 14(1):37–85, 2004.
14. G. Montgomery. Color blindness: More prevalent among males, 1997. http://www.hhmi.org/senses/b130.html.
15. A. J. Munro, K.Höök, and D. Benyon. *Social navigation of information space*. Springer, 1999.
16. R. Picard. *Affective Computing*. MIT Press, 1997.
17. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *CSCW '94: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pages 175–186. ACM Press, 1994.
18. P. Resnick and H. R. Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, 1997.
19. A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: Toward evaluation studies of tagclouds. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 995–998. ACM, 2007.
20. J. A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273 – 294, 1977.
21. J. Schrammel, M. Leitner, and M. Tscheligi. Semantically structured tag clouds: An empirical evaluation of clustered presentation approaches. In *CHI '09: Proceedings of the 27th International Conference on Human Factors in Computing Systems*, pages 2037–2040. ACM, 2009.
22. S. Sen, F. M. Harper, A. LaPitz, and J. Riedl. The quest for quality tags. In *GROUP '07: Proceedings of the 2007 International ACM Conference on Supporting Group Work*, pages 361–370. ACM, 2007.
23. 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 *CSCW '06: Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, pages 181–190. ACM, 2006.
24. S. Sen, J. Vig, and J. Riedl. Learning to recognize valuable tags. In *IUI '09: Proceedings of the 13th International Conference on Intelligent User Interfaces*, pages 87–96. ACM, 2009.
25. S. Sen, J. Vig, and J. Riedl. Tagommenders: Connecting users to items through tags. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web*, pages 671–680. ACM, 2009.
26. J. Vig, S. Sen, and J. Riedl. Tagsplanations: Explaining recommendations using tags. In *IUI '09: Proceedings of the 13th International Conference on Intelligent User Interfaces*, pages 47–56. ACM, 2009.
27. R. Wetzker, W. Umbrath, and A. Said. A hybrid approach to item recommendation in folksonomies. In *ESAIR '09: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 25–29. ACM, 2009.

[15] http://www.citeulike.org